READING THE WHOLE HEART: LATENT-ATTENTION MASKED AUTOENCODERS FOR MULTIMODAL CARDIAC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiovascular diagnosis and treatment rest on integrating complementary modalities, such as electrocardiogram, echocardiography, and chest X-rays, each capturing distinct but complementary aspects of cardiac pathophysiology. Yet most medical foundation models remain modality-specific, combining modalities only for finetuning or post-training. This discards the cross-modal evidence clinicians naturally integrate and ignores the structure within each modality. We introduce Latent-Attention Masked Autoencoder (LAMAE), a multimodal, structure-aware masked autoencoder that jointly learns patient-level representations during self-supervised pretraining. Instead of fusing modalities post hoc, LAMAE exchanges information directly in the latent space through a shared latent-attention module operating over a study–view–entity hierarchy, enabling aggregation of variable observations and handling of missing modalities. Pretrained on over $500{,}000$ MIMIC-IV hospital stays, LAMAE outperforms modality-specific pretraining and strong contrastive and vision–language baselines across multimodal hospital-stay tasks, such as in-hospital mortality, ICD-10 and DRG coding, and length of stay, while remaining competitive on unimodal tasks. Modeling this structure also pays off within a single modality: even without cross-modal information, the latent-attention module improves representations over modality-specific pretraining.

1 INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for more than 20 million deaths annually and imposing a substantial medical and financial burden on healthcare systems (Roth et al., 2025). Thus, early diagnosis and appropriate risk stratification are critical for improving patient outcomes. Notably, clinical decision making rarely relies on a single diagnostic modality (Acosta et al., 2022). Instead, clinicians routinely integrate complementary information from electrocardiograms (ECGs), echocardiography (Echo), chest radiographs (CXR), laboratory values, and clinical history to form a comprehensive assessment of cardiovascular diagnosis and prognosis (Kashou et al., 2023; Dohi, 2019).

Recent advances in self-supervised learning (SSL) and foundation models have demonstrated that large-scale pretraining can learn transferable representations from unlabeled medical data, thereby substantially reducing the dependence on expensive expert annotations (Azizi et al., 2021; Khan et al., 2025). Masked autoencoders (MAEs, He et al., 2022) have emerged as a scalable paradigm for representation learning by reconstructing masked portions of the input from sparse observations. MAEs have since been successfully adapted to numerous medical domains, including CXRs (Xiao et al., 2023), Echocardiography (Jiao et al., 2024; Kim et al., 2025), and ECG (Na et al., 2024; Jin et al., 2025). Despite these advances, nearly all existing foundation models, and especially their pretraining strategies, remain modality-specific, learning separate representations for ECG, Echo, or CXR independently before combining them. If at all, different modalities are only combined during downstream prediction.

Beyond being modality-specific, many existing approaches also overlook the intrinsic structure present within medical data. Medical observations are rarely independent. A standard 12-lead ECG consists of multiple complementary voltage vector projections of the same cardiac electrical

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.

activity, while CXR and Echo examinations comprise multiple anatomical views acquired from different imaging angles. These findings suggest that exploiting structural relationships constitutes an effective form of inductive bias for medical foundation models. However, current approaches remain confined to individual modalities and therefore fail to exploit the relationships between complementary diagnostic modalities acquired from the same patient. In current clinical practice, clinicians take advantage of these diagnostic modalities for many clinically important diseases, including acute myocardial infarction, heart failure, valvular heart disease, and cardiomyopathies (McDonagh et al., 2021; Byrne et al., 2023). Each modality provides complementary evidence, which clinicians naturally integrate for diagnosis.

However, learning from heterogeneous diagnostic modalities presents several challenges: First, different modalities exhibit different data characteristics, ranging from one-dimensional physiological time series to two-dimensional images and spatio-temporal videos. Second, clinical examinations are inherently incomplete: patients often undergo only a subset of available diagnostic procedures, requiring models to handle arbitrary patterns of missing modalities. Finally, multimodal observations exhibit hierarchical structure, with dependencies both within modalities (e.g., ECG leads or echocardiographic views) and across modalities at the patient level. Existing multimodal learning approaches typically rely on contrastive alignment or late feature fusion, limiting their ability to directly model these hierarchical interactions during representation learning.

In this work, we propose Latent-Attention Masked Autoencoder (LAMAE), a multimodal and structure-aware MAE (He et al., 2022) that jointly learns representations from ECG, Echo, and CXRs through a shared latent attention (LA) mechanism. Rather than treating modalities as independent feature extractors whose representations are fused only during downstream prediction, our LAMAE enables information exchange directly within the latent space during self-supervised pretraining. This design naturally accommodates heterogeneous modalities, variable numbers of observations, and missing data while preserving the flexibility to exploit modality-specific information when appropriate. Our framework learns unified patient-level representations that better reflect the multimodal nature of clinical reasoning.

Our contributions are three-fold: (1) We introduce Latent-Attention Masked Autoencoders (LAMAE), a structure-aware multimodal masked autoencoder for ECG, Echo, and CXRs: a shared latent-attention (LA) module exchanges information directly in the latent space over a study–view–entity hierarchy, paired with an entity-level masking scheme that drops complete leads, frames, or views to force reconstruction across modalities and views, as illustrated in Figure 1. (2) We show that modeling this structure improves representations at every level: LAMAE outperforms modality-specific pretraining and strong contrastive and vision–language baselines on hospital-stay-level tasks, and the LA module improves representations even within a single modality, without any cross-modal information. (3) We show that multimodal pretraining outperforms unimodal pretraining on hospital-stay tasks without sacrificing unimodal performance.

2 Related Work

Self-supervised pretraining in medicine. Self-supervised learning has become a key strategy for medical representation learning, where expert annotation is costly and labels are imbalanced (Azizi et al., 2021; Moody et al., 2025). Contrastive frameworks (Chen et al., 2020; He et al., 2020) applied to large corpora such as MIMIC-CXR (Johnson et al., 2019) yield encoders that transfer well (Azizi et al., 2021; Tiu et al., 2022). However, they treat every image as an independent instance. Masked autoencoders (MAEs) reconstruct masked patches (He et al., 2022; Bizeul et al., 2025) or spatiotemporal tubes (Tong et al., 2022; Wang et al., 2023). They have been adapted separately to CXR (Xiao et al., 2023), Echo (Jiao et al., 2024; Kim et al., 2025; Zhang et al., 2024; Yang et al., 2026) and ECG (Na et al., 2024; Jin et al., 2025). The resulting encoders are siloed, and their reconstruction signal never leaves a single modality.

Multimodal medical models. Vision–language models pair images with reports (Boecking et al., 2022; Chen et al., 2024; Deperrois et al., 2025), generalist models extend this to several imaging domains (Sellergren et al., 2026; Zhang et al., 2026; Gao et al., 2025), and ImageBind (Girdhar et al., 2023) aligns many modalities to a single anchor. In cardiology, Echo2ECG (Liman et al., 2026) contrastively aligns ECGs with multi-view echocardiograms. Text-anchored supervision is semantic rather than physiological: a report states what is seen, not how the different recordings of

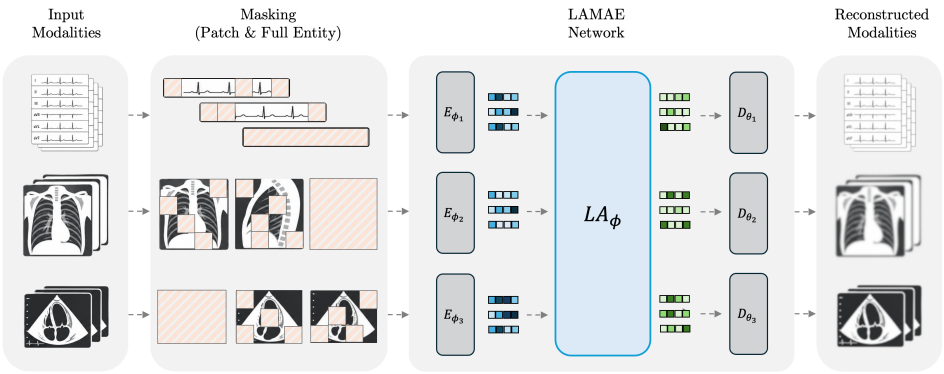


Figure 1: **Overview of the LAMAE framework.** From left to right: *(1) Input modalities:* a sample contains a variable set of ECG, CXR, and Echo observations. *(2) Masking:* inputs are masked at both the patch and full-entity levels, with entity masking removing complete entities such as ECG leads, CXR views, or Echo frames. *(3) LAMAE network:* modality-specific encoders E_{ϕ_k} independently encode the visible tokens, after which the shared latent attention module LA_ϕ jointly processes their latent representations, enabling information exchange within and across modalities. Modality-specific decoders D_{θ_k} then reconstruct the masked inputs. *(4) Reconstructed modalities:* the reconstructed ECG, CXR, and Echo observations provide the self-supervised pretraining objective.

the same heart relate to each other. Contrastive alignment, whether to text or between modalities, additionally requires paired samples, which are scarce for many modality combinations in routine care (fig. 4).

Multimodal masked autoencoders. MultiMAE (Bachmann et al., 2022) and 4M (Mizrahi et al., 2023) reconstruct masked tokens of several modalities with a single shared encoder, and M3AE (Geng et al., 2022) does the same for image–text pairs. Turgut et al. (2025) combine masked data modeling with contrastive alignment to transfer information from cardiac MRI to the ECG, but this requires paired studies of two modalities. LAMAE instead keeps modality-specific encoders at the entity level and fuses modalities only through a shared latent-attention module. This lets it pretrain on any observed subset of modalities.

Structured studies. Medical data are rarely single-entity samples: a CXR study couples frontal and lateral projections, an ECG comprises twelve simultaneous leads, and an echocardiographic exam consists of several videos from complementary views. Prior work exploits this through multi-view or correspondence-based contrast (Sutter et al., 2024; Agostini et al., 2025; Laguna et al., 2025), hierarchical fusion (Mokhtari et al., 2023), latent-space MAEs over frozen per-view encoders (Tohyama et al., 2025), or lead alignment combined with masked reconstruction (Erlacher et al., 2025). These approaches mostly rely on *pairwise* objectives. LAMAE instead provides a unified framework for modeling structure both within and across modalities, jointly integrating studies, views, and entities through latent attention. In this way, LAMAE learns not from *more* data, but from the *structure* of the data itself.

3 Method

We propose Latent-Attention Masked Autoencoders (LAMAEs), a multimodal and structure-aware MAE (He et al., 2022) that jointly learns representations from multimodal datasets $\mathbb{S}$ (Figure 1). We define a multimodal dataset as $\mathbb{S} = \{\mathcal{S}^{(s)}\}_{s=1}^{S}$ where every multimodal sample $\mathcal{S}^{(s)} = \left(\{\boldsymbol{X}^{(s,k)}\}_{k\in\mathcal{M}}, \boldsymbol{m}^{(s)}\right)$ pairs the per-modality data $\boldsymbol{X}^{(s,k)}$, $k \in \mathcal{M}$, with an availability mask $\boldsymbol{m}^{(s)} \in \{0,1\}^{|\mathcal{M}|}$ whose entries are

$$m^{(s,k)} = \begin{cases} 1, & \text{if modality } k \text{ is observed for } s, \\ 0, & \text{otherwise.} \end{cases} . \tag{1}$$

The observed modalities per multimodal sample s are then $\mathcal{M}_{\text{obs}}^{(s)} = \{k \in \mathcal{M} \mid m^{(s,k)} = 1\}$.

The data acquisition process of healthcare data leads to more structure in the data compared to general or natural domain data (Laguna et al., 2025). Therefore, we introduce a three-layered hierarchy to describe modalities: *studies*, *views*, *and base entities*. Every modality $\boldsymbol{X}^{(s,k)}$ is a medical study that may consist of multiple examinations, e.g., an Echo study usually consists of multiple videos of different frames each. Hence, a modality is given as $\boldsymbol{X}^{(s,k)} = \{X^{(s,k,1)}, \ldots, X^{(s,k,V_k)}\}$, where $X^{(s,k,v)}$ is the data sample for view v and modality k, e.g., video of an echo study. For every modality k and sample s, we define the set of observed views as $\mathcal{V}_{k,\text{obs}}^{(s)} \subseteq \mathcal{V}_k = \{1, \ldots, V_k\}$, where $V_k = |\mathcal{V}_k|$ is the total number of views and $\mathcal{V}_k$ the complete set of views for modality k. Every view v of modality k is defined as $X^{(s,k,v)} = (\boldsymbol{x}^{(s,k,v,1)}, \ldots, \boldsymbol{x}^{(s,k,v,E_k)})$, where each base entity $\boldsymbol{x}^{(s,k,v,e)}$ is, e.g., a frame in an Echo video, a lead in an ECG measurement, or an image in a CXR study, and $\mathcal{E}_k = \{1, \ldots, E_k\}$ is the set of base entities with $E_k = |\mathcal{E}_k|$. We illustrate this hierarchy for an ECG study in Figure 2.

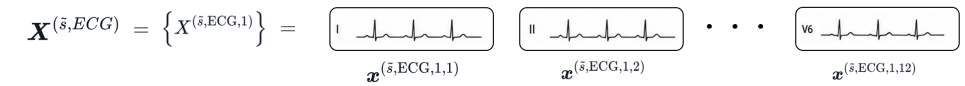


Figure 2: **Example of the study-view-entity hierarchy for ECG.** For an example sample $\tilde{s}$, the ECG modality $\boldsymbol{X}^{(\tilde{s},\text{ECG})}$ consists of a single ECG examination $X^{(\tilde{s},\text{ECG},1)}$ ($v = 1$), which comprises $E_{\text{ECG}} = 12$ base entities $\boldsymbol{x}^{(\tilde{s},\text{ECG},1,e)}$, corresponding to the 12 ECG leads. Leads I, II, and V6 are shown for illustration.

3.1 Encoders and Decoders

We have modality-specific encoders $E_{\phi_k}(\cdot)$ and decoders $D_{\theta_k}(\cdot)$. Following He et al. (2022), we use transformer-based encoder and decoder architectures (Vaswani et al., 2017). More specifically we follow the suggested model variants presented in Dosovitskiy et al. (2021) and He et al. (2022). For details regarding the modality-specific adaptations, we refer to section A.

As it is typical for MAEs, we mask a large portion of the input $\boldsymbol{X}^{(s,k)}$ during pretraining with the objective to reconstruct the masked input. Given a binary mask $M_{k,\alpha_k}^{(s)}(\cdot)$ with mask ratio α_k for modality k, we define the model input $\boldsymbol{X}_{vis}^{(s,k)}$ as

$$\boldsymbol{X}_{vis}^{(s,k)} = M_{k,\alpha_k}^{(s)}(\boldsymbol{X}^{(s,k)}). \tag{2}$$

During pretraining, the encoder E_{ϕ_k} only processes the visible part $\boldsymbol{X}_{vis}^{(s,k)}$ and embeds them into latent embeddings $\boldsymbol{Z}^{(s,k)}$ to reconstruct the input $\hat{\boldsymbol{X}}^{(s,k)}$. For every modality k, the encoder E_{ϕ_k} sequentially processes modalities on a base entity level $\boldsymbol{x}^{(s,k,v,e)}$. For every $\boldsymbol{x}^{(s,k,v,e)}$, we have T_k patches and $T_{k,vis} = (1 - \alpha_k) \cdot T_k$ visible patches. Hence, $\mathcal{T}_{k,vis}^{(s)} \subseteq \{1, \ldots, T_k\}$ is the set of input patches for entity e of modality k. We then have

$$\boldsymbol{z}^{(s,k,v,e)} = E_{\phi_k}(\boldsymbol{x}^{(s,k,v,e)}) \quad \text{such that} \quad \boldsymbol{Z}^{(s,k)} = \{\boldsymbol{z}^{(s,k,v,e)} \mid e \in \mathcal{E}_k, v \in \mathcal{V}_{k,\text{obs}}^{(s)}\} \tag{3}$$

Following standard practice in ViTs (Dosovitskiy et al., 2021), each latent embedding $\boldsymbol{z}^{(s,k,v,e)}$ consists of the latent embeddings $\boldsymbol{z}_t^{(s,k,v,e)}$ of the visible patches $t \in \mathcal{T}_{k,vis}^{(s)}$ and an entity-level CLS embedding $\boldsymbol{z}_{\text{cls}}^{(s,k,v,e)}$. For simplicity, we describe the set of latent embeddings $\boldsymbol{Z}^{(s,k)}$ for modality k as $\boldsymbol{Z}^{(s,k)} = E_{\phi_k}(\boldsymbol{X}_{vis}^{(s,k)})$. Similarly, we have modality-specific decoders D_{θ_k}, which decode the base entities of each modality separately, i.e.,

$$\hat{\boldsymbol{x}}^{(s,k,v,e)} = D_{\theta_k}(\boldsymbol{z}_{out}^{(s,k,v,e)}) \quad \text{where} \quad \boldsymbol{Z}_{out}^{(s,k)} = \{\boldsymbol{z}_{out}^{(s,k,v,e)} \mid e \in \mathcal{E}_k, v \in \mathcal{V}_{k,\text{obs}}^{(s)}\}. \tag{4}$$

For vanilla MAEs, we have $\boldsymbol{z}^{(s,k,v,e)} = \boldsymbol{z}_{out}^{(s,k,v,e)}$.

3.2 Latent Attention for Multimodal Integration

Multimodal datasets $\mathbb{S}$ provide additional structure in the data acquisition process that we can leverage. We introduce a novel *Latent Attention* (LA) module for leveraging the additional structure in the

data. Inspired by Ilse et al. (2018) and Lee et al. (2019), the LA module allows to learn the correlation and shared information between different modalities, views or entities. The LA module is where the information coming from all modalities is being processed jointly. Let $\boldsymbol{Z}^{(s)} = \{\boldsymbol{Z}^{(s,k)}\}_{k \in \mathcal{M}_{\text{obs}}^{(s)}}$ be the concatenation of latent embeddings $\boldsymbol{Z}^{(s,k)}$ of all modalities $k \in \mathcal{M}_{\text{obs}}^{(s)}$. Given a set of latent embeddings $\boldsymbol{Z}^{(s)}$, the LA module transforms the embeddings $\boldsymbol{z}^{(s,k,v,e)}$ that only contain information of their respective base entities into embeddings containing all available multimodal information as follows

$$\boldsymbol{Z}_{out}^{(s)} = LA_{\phi}(\boldsymbol{Z}^{(s)}) = \{\boldsymbol{z}_{out}^{(s,k,v,e)} \mid k \in \mathcal{M}_{\text{obs}}^{(s)}, v \in \mathcal{V}_{k,\text{obs}}^{(s)}, e \in \mathcal{E}_k\}. \tag{5}$$

We design the LA module as a multi-head self-attention block that attends jointly over the patch and CLS embeddings of all observed entities, views and modalities (see Section A).

3.3 Entity Masking

The LA module allows for more elaborate masking strategies compared to single-entity-based approaches, such as the independent (or vanilla) MAE (He et al., 2022). Independent MAEs define their pretraining objective via patch-based masking: a signal or image is divided into patches, of which a random subset is masked, defining the reconstruction target (see section 3.4 for details). The LA module additionally allows us to mask or drop complete entities such as leads, frames, or views. Dropping complete entities forces the LA module to connect information across base entities.

We make Equation (2) more specific as follows. Let $\mathcal{E}_k$ denote the entity set of modality k, with $E_k = |\mathcal{E}_k|$ assumed constant across the views v of modality k, and assume all entities of a given (k, v) carry the same number of tokens. For each sample s, modality k, and view v, we independently sample the number of visible (encoded) entities

$$\mathcal{E}_{vis}^{(s,k,v)} \subseteq \mathcal{E}_k \quad \text{with} \quad \left|\mathcal{E}_{vis}^{(s,k,v)}\right| = n_{e,vis}^{(s,k,v)} \sim \mathcal{U}\{n_{\min}, n_{\max}\}, \tag{6}$$

and treat the remaining $E_k - n_{e,vis}^{(s,k,v)}$ dropped entities as fully masked (mask ratio 100%). The reconstruction target comprises all masked tokens, i.e., every token of a dropped entity together with the masked tokens of the visible entities.

We target a mask ratio α_k per modality, held constant across all views v and all samples s. Since $n_{e,vis}^{(s,k,v)}$ varies across views, we set the per-entity mask ratio $\beta^{(s,k,v)}$ applied to the visible entities so that each view attains the common target α_k in expectation. Requiring $\big(E_k - n_{e,vis}^{(s,k,v)} + n_{e,vis}^{(s,k,v)}\beta^{(s,k,v)}\big)/E_k = \alpha_k$ gives

$$\beta^{(s,k,v)} = \frac{\alpha_k - 1 + n_{e,vis}^{(s,k,v)}/E_k}{n_{e,vis}^{(s,k,v)}/E_k}. \tag{7}$$

For $\beta^{(s,k,v)} \geq 0$ the number of dropped entities must not by itself exceed the target, which bounds the sampler via $n_{\min} \geq \lceil (1 - \alpha_k)\, E_k \rceil$.

The resulting masking procedure is

$$\boldsymbol{X}_{vis}^{(s,k)} = \left\{\, M_{k,\beta^{(s,k,v)}}(\boldsymbol{x}^{(s,k,v,e)}) \,\right\}_{v,\, e \in \mathcal{E}_{vis}^{(s,k,v)}} = M_{k,\alpha_k}^{(s)}\big(\boldsymbol{X}^{(s,k)}\big), \tag{8}$$

i.e. the per-view entity-dropping combined with the per-view ratio $\beta^{(s,k,v)}$ is, in expectation and under the equal-token assumption within each (k, v), equivalent to applying a single mask ratio α_k to modality k.

3.4 LAMAE

During pretraining, we optimize the following objective $\mathcal{L}^{(s)}$ for every hospital stay s

$$\mathcal{L}^{(s)} = \sum_{k \in \mathcal{M}_{\text{obs}}^{(s)}} \frac{1}{|\mathcal{V}_{k,\text{obs}}^{(s)}|} \sum_{v \in \mathcal{V}_{k,\text{obs}}^{(s)}} \frac{1}{E_k} \sum_{e \in \mathcal{E}_k} \frac{1}{\alpha_k T_k} \sum_{t \notin \mathcal{T}_{k,vis}^{(s)}} \left\| \boldsymbol{x}_t^{(s,k,v,e)} - \hat{\boldsymbol{x}}_t^{(s,k,v,e)} \right\|_2^2, \tag{9}$$

where $\hat{\boldsymbol{x}}^{(s,k,v,e)} = D_{\theta_k}\big(\big[LA_\phi(\boldsymbol{Z}^{(s)})\big]^{(k,v,e)}\big)$ and $\big[LA_\phi(\boldsymbol{Z}^{(s)})\big]^{(k,v,e)} = \boldsymbol{z}_{out}^{(s,k,v,e)}$ denotes the output embeddings of base entity e in view v of modality k. A key property of our design is that it handles missing modalities by construction, rather than requiring complete cases or imputation as many multimodal methods do. Because the reconstruction objective is applied only over the observed modalities, and the latent-attention (LA) module aggregates a variable-size set of available modality tokens, the model both trains and performs inference on any subset of modalities without architectural change or retraining. Because every output embedding can attend to all observed modalities, each entity-level CLS embedding after the LA module carries information from the whole hospital stay.

For downstream tasks, we represent a hospital stay by the average of the entity-level CLS embeddings after the LA module,

$$\bar{\boldsymbol{z}}^{(s)} = \frac{1}{N^{(s)}} \sum_{k \in \mathcal{M}_{\text{obs}}^{(s)}} \sum_{v \in \mathcal{V}_{k,\text{obs}}^{(s)}} \sum_{e \in \mathcal{E}_k} \boldsymbol{z}_{\text{cls},out}^{(s,k,v,e)}, \tag{10}$$

where $\boldsymbol{z}_{\text{cls},out}^{(s,k,v,e)}$ is the CLS embedding of base entity e in $\boldsymbol{Z}_{out}^{(s)}$ and $N^{(s)}$ is the number of observed base entities of stay s.

4 Experiments & Results

Our aim is to evaluate how the LA block influences the pretraining. We do so by assessing the downstream task performance of various tasks, both on a hospital-stay level (see Section 4.2), where we expect multimodal models to benefit, and on unimodal tasks, where unimodal models should have all the required information to solve the respective unimodal tasks (see Section 4.3). These two settings reflect the different layers of tasks within the healthcare setting, which range from low-level modality-specific tasks such as left ventricular ejection fraction estimation (LVEF) to tasks that reflect the longitudinal nature of a patient journey within a hospital stay such as in-hospital mortality or diagnosis codes. We denote by LAMAE_{MM} the model pretrained on all three modalities, and by LAMAE_k and Independent MAE_k, $k \in \{\text{CXR}, \text{ECG}, \text{Echo}\}$, models pretrained on a single modality k.

We study four complementary questions:

(i) *Does cross-modal interaction during pretraining improve hospital-stay representations?*
LAMAE_{MM} outperforms Independent MAE_{MM} on all six hospital-stay tasks and both state-of-the-art baselines on five of six tasks (Table 1).

(ii) *Does multimodal pretraining outperform unimodal pretraining on hospital-stay tasks?*
LAMAE_{MM} outperforms every unimodal LAMAE_k on the complete test set, demonstrating its advantage in clinically realistic settings with heterogeneous modality availability (Table 2). Importantly, LAMAE_{MM} also outperforms most of the unimodal models on a modality constrained test set across modality configurations, indicating that its gains extend beyond increased coverage and reflect the benefit of the multimodal architecture itself (Table 3).

(iii) *Does the LA module improve representations within a single modality?*
LAMAE_k outperforms Independent MAE_k on CXR, Echo, and most ECG targets (Tables 3a to 3c).

(iv) *Does multimodal pretraining come at the expense of unimodal performance?*
No. Finetuned on a single modality, LAMAE_{MM} matches or exceeds LAMAE_k on most unimodal tasks (Tables 3a to 3c).

4.1 Data

We use the MIMIC-IV database (Johnson et al., 2023; Goldberger et al., 2000), together with its modality-specific subsets MIMIC-CXR (Johnson et al., 2019), MIMIC-IV-ECG (Gow et al., 2023), and MIMIC-IV-Echo (Gow et al., 2026) to generate a dataset of multimodal hospital stays. A hospital stay is a collection of ordered modalities describing the health status of a patient. Note that not every hospital stay $\mathcal{S}^{(s)}$ is comprised of the same modalities as every patient trajectory potentially requires a different set of modalities to be acquired. In total, our dataset consists of 658,847 unique

Table 1: **Hospital Stay Tasks.** We finetune all methods using 50,000 labeled samples. Every sample has at least one ECG, Echo, or CXR measurement. At test, the model receives the set of modalities that is available for each hospital stay. We report the AUROC (%) or macro average of AUROC where applicable (mean ± std).

	Mortality	ICD10	DRG			Length of Stay
Model	*In-hospital*	*Ch. IX*	*Code*	*Severity*	*Mortality Risk*	*In-hospital*
MedSigLip	89.62±0.38	82.06±0.18	**59.39**±0.38	68.75±0.22	70.65±0.06	79.20±0.44
ProbMED	90.20±0.66	82.56±0.08	59.38±0.35	69.31±0.28	71.36±0.34	79.81±0.38
Independent MAE_{MM}	90.51±0.47	81.66±0.11	57.46±0.23	69.08±0.36	71.14±0.23	80.87±0.25
$LAMAE_{MM}$	**91.59**±0.37	**83.02**±0.18	58.92±0.31	**70.29**±0.24	**72.21**±0.43	**81.70**±0.26

hospital stays with a total of 227,835 CXR studies and 377,110 images, 7,243 Echo studies and 525,328 videos, and 800,035 12-lead ECG measurements.

4.2 Multimodal Evaluation

Clinical questions at the level of a hospital stay, such as diagnosis, length of stay or mortality risk, depend on integrating imaging, waveform and clinical information collected during the stay rather than on any single measurement. We therefore evaluate LAMAE on hospital-stay-level tasks, where this kind of cross-modal reasoning is most directly required.

Setup. We evaluate on six hospital-stay-level tasks from four task groups: in-hospital mortality, ICD-10 Chapter IX diagnosis coding (World Health Organization, 2004), DRG classification (code, severity and mortality risk) (Fetter et al., 1980), and length of stay. We compare against two state-of-the-art baselines, ProbMED (Gao et al., 2025) and MedSigLip (Sellergren et al., 2026) (details in Section D). In addition, we include an Independent MAE that is trained on all three modalities with the same data and training budget as LAMAE, but without the LA module and the entity masking (Table 5); its modalities therefore never interact during pretraining. To test whether multimodal pretraining and multimodal input are beneficial, we further pretrain and finetune one LAMAE per modality ($LAMAE_{CXR}$, $LAMAE_{ECG}$, $LAMAE_{Echo}$). Each of these unimodal models can score only the stays that contain its modality; all other test stays receive the training-set prior as prediction, so that every model is evaluated on the same test set (Section D). All models are finetuned on 50,000 labeled hospital stays using the same finetuning configuration (Table 7). We select the checkpoint with the lowest validation loss and report mean and standard deviation over three seeds on the test set.

Results. Table 1 shows that $LAMAE_{MM}$ achieves the best AUROC on five of the six tasks, outperforming ProbMED, MedSigLip and the Independent MAE on in-hospital mortality, ICD-10, DRG severity, DRG mortality risk and length of stay. On DRG code prediction, the two baselines are slightly ahead (59.39 and 59.38 vs. 58.92), with differences of about one standard deviation. The comparison with the Independent MAE highlights the contribution of cross-modal interaction during pretraining: LAMAE improves on every task, by 0.8 to 1.5 percentage points. The Independent MAE on its own is already on par with the state-of-the-art baselines, ahead on mortality and length of stay and behind on ICD-10 and DRG code, which suggests that LAMAE's advantage stems from modeling cross-modal interactions rather than from the multimodal pretraining data alone. Table 2 compares $LAMAE_{MM}$ with LAMAE models pretrained and finetuned on a single modality. $LAMAE_{MM}$ outperforms all unimodal models on every task. The strongest unimodal model, $LAMAE_{ECG}$, is based on the most widely available modality, but is still behind $LAMAE_{MM}$ by 2.7 to 7.4 percentage points, with the largest gaps on ICD-10 and length of stay. $LAMAE_{CXR}$ and especially $LAMAE_{Echo}$ remain close to chance, because they can score only the small fraction of stays that contain their modality. These gaps therefore combine two effects: the information contributed by the other modalities, and the ability of $LAMAE_{MM}$ to make a prediction for any stay with at least one observed modality. Removing cross-modal interaction (Independent MAE) and removing modalities (unimodal LAMAEs) both degrade performance, showing that multimodal pretraining with the LA module and access to multiple modalities are both needed for the best hospital-stay performance.

Table 2: **Unimodal vs. Multimodal Models.** The subscript gives the set of modalities the model was pretrained on and takes as input. Test samples that lack an accepted modality are assigned the training-set prior, so every row is scored on the same test set. We report AUROC (%), or the macro average of AUROC where applicable (mean ± std).

	Mortality	ICD10	DRG			Length of Stay
Model	*In-hospital*	*Ch. IX*	*Code*	*Severity*	*Mortality Risk*	*In-hospital*
$\text{LAMAE}_{\text{Echo}}$	49.80±0.24	50.45±0.09	42.36±0.05	50.20±0.14	50.31±0.09	50.47±0.03
$\text{LAMAE}_{\text{CXR}}$	61.03±0.20	56.53±0.05	51.25±0.11	52.91±0.16	53.35±0.06	57.48±0.04
$\text{LAMAE}_{\text{ECG}}$	86.22±0.69	75.67±0.52	55.65±0.68	67.50±0.11	69.56±0.08	74.47±0.21
LAMAE_{MM}	**91.59**±0.37	**83.02**±0.18	**58.92**±0.31	**70.29**±0.24	**72.21**±0.43	**81.70**±0.26

4.3 Unimodal Evaluation

While we propose LAMAE mainly for its multimodal capabilities, its hierarchical approach can be leveraged equally well for unimodal problems and datasets where some structure is present to be leveraged. Therefore, we evaluate LAMAE on unimodal tasks too, validating the unimodal performance of the proposed LAMAE approach. Additionally, we report ICD-10 classification performance, but unlike in Table 2, we evaluate only on records where the examined modality is present, isolating how well the model uses that modality. For all modalities, we select the epoch with the lowest validation loss and report test performance over three seeds.

Table 3: **Unimodal Results.** We evaluate the capabilities of the latent attention module to extract meaningful representations also for unimodal data. We compare a LAMAE that is trained only on a single data type to an independent MAE that is not leveraging any structure in the data. We report finetuning results on the test set (mean ± std over 3 seeds). We compare the models on relevant unimodal tasks and the classification of ICD10 codes (see Table 1).

(a) **CXR.** We use the 14 CheXpert labels (Irvin et al., 2019) to assess the understanding of CXR data and report the macro AUROC (%, ↑).

Method	CXR Tasks	ICD10
Independent MAE_{CXR}	82.58±0.07	85.03±0.05
$\text{LAMAE}_{\text{CXR}}$	83.29±0.04	85.56±0.13
LAMAE_{MM}	**83.42**±0.05	**85.88**±0.13

(b) **Echo.** We evaluate the Echo-specific performance in the left ventricular ejection fraction (LVEF) regression task, where we report the R^2 score (↑).

Method	LVEF	ICD10
Independent MAE_{Echo}	0.273±0.008	57.49±0.40
$\text{LAMAE}_{\text{Echo}}$	**0.414**±0.055	**59.21**±0.71
LAMAE_{MM}	**0.414**±0.009	59.10±0.40

(c) **ECG.** We evaluate the ECG interval regression, reported in R^2 score (↑), alongside the ECG-specific ICD-10 task AUROC (%) (↑).

Method	RR	PR	P wave	QRS	QT	QTc	ICD10
Independent MAE_{ECG}	0.91±0.00	0.13±0.05	−0.36±0.28	0.51±0.17	0.58±0.07	0.23±0.05	70.08±1.71
$\text{LAMAE}_{\text{ECG}}$	0.95±0.01	0.43±0.02	−0.03±0.06	**0.53**±0.33	0.81±0.02	**0.54**±0.05	76.87±0.52
LAMAE_{MM}	**0.98**±0.01	**0.67**±0.03	**−0.02**±0.08	0.51±0.10	**0.85**±0.05	0.41±0.32	**79.13**±0.30

Chest X-Ray

Setup. We evaluate on the 14-label CheXpert finding classification task described above, with each CXR study capped at a maximum of three views. Alongside the view-based Independent MAE_{CXR} baseline, trained solely on CXR, we evaluate $\text{LAMAE}_{\text{CXR}}$, and additionally report LAMAE_{MM}, the same fully multimodal-pretrained checkpoint used in the Echo evaluation, finetuned and tested with its input restricted to CXR only. Pretraining uses all available CXR studies for 1000 epochs; finetuning runs for 50 epochs.

Result. Table 3a reports the results. $\text{LAMAE}_{\text{CXR}}$ improves over the Independent MAE_{CXR} on both CXR findings and ICD-10 classification, increasing macro AUROC from 82.58 to 83.29 and from 85.03 to 85.56, respectively. LAMAE_{MM} further improves performance to 83.42 and 85.88, despite being evaluated with CXR input only. This indicates that multimodal pretraining improves the CXR

representation even when no additional modalities are available at finetuning or test time. Figure 7 shows how this aggregate score breaks down across the 14 individual CheXpert findings.

Echocardiogram

Setup. We predict left ventricular ejection fraction (LVEF), taken from the structured measurements accompanying the MIMIC-IV-Echo studies (Gow et al., 2026). Alongside the frame-based Independent MAE_{Echo} baseline, we evaluate $\text{LAMAE}_{\text{Echo}}$ and additionally report LAMAE_{MM}, the same multimodal-pretrained checkpoint used in the evaluation of section 4.2, finetuned and tested with its input restricted to Echo only. From each video (selection described in Section D), we sample 8 frames with a stride of 4 at a random offset. Hyperparameters are listed in Tables 5 and 6.

Result. Table 3b reports the results. Both LAMAE variants substantially improve over the baseline, raising LVEF R^2 from 0.273 to 0.414 and ICD-10 AUROC from 57.49 to 59.21 and 59.10, with the two variants within seed variance of each other.

Electrocardiogram

Setup. We derive six global interval measurements from the MIMIC-IV-ECG machine annotations (Strodthoff et al., 2024): the RR interval, taken directly from the annotations, and the PR interval, P-wave duration, QRS duration and QT interval, obtained as differences of the annotated wave onsets and offsets. PR is the interval from P onset to QRS onset, P wave from P onset to P end, QRS from QRS onset to QRS end, and QT from QRS onset to T end. The rate-corrected QTc interval is computed from QT and RR using Bazett's formula, following Plagwitz et al. (2025). Alongside the lead-based Independent MAE_{ECG} baseline, we evaluate $\text{LAMAE}_{\text{ECG}}$ and additionally report LAMAE_{MM}, the same multimodal-pretrained checkpoint used in the evaluation of section 4.2, finetuned and tested with its input restricted to ECG only. Hyperparameters are listed in Tables 5 and 6.

Result. Table 3c and Section D report the results. Both LAMAE variants substantially improve over the baseline on PR, QT and QTc, and on ICD-10 classification (70.08 to 76.87 and 79.13 AUROC), with LAMAE_{MM} performing best on most targets. QRS duration is within seed variance across all models, and P-wave duration is not recovered by any model ($R^2 \leq 0$).

Summary. Overall, LAMAE improves representation learning even in the unimodal setting, showing that its benefits extend beyond multimodal integration through its ability to naturally leverage the inherent structure of medical data.

5 Conclusion

We introduced Latent-Attention Masked Autoencoders (LAMAE), a multimodal, structure-aware masked autoencoder that learns patient-level representations from ECG, Echo, and CXR during self-supervised pretraining. Instead of fusing modalities only at the prediction stage, LAMAE exchanges information across modalities in the latent space already during pretraining. It does so through a shared latent-attention (LA) module operating over a study–view–entity hierarchy, naturally handling heterogeneous data, variable numbers of observations, and missing modalities within a single reconstruction objective. Pretrained on more than $500{,}000$ MIMIC-IV hospital stays, LAMAE outperforms an otherwise identical model without cross-modal interaction, models pretrained on a single modality, and strong contrastive and vision–language baselines on hospital-stay-level tasks. The benefits of modeling structure extend to the unimodal setting: the LA module improves over independent MAEs on CXR, Echo, and ECG tasks, and the multimodal-pretrained model matches or exceeds its modality-specific counterparts when finetuned on a single modality, indicating that cross-modal pretraining does not come at the expense of unimodal performance. This work has limitations: all experiments use data from a single institution, Echo is available for only a small fraction of hospital stays, and some fine-grained targets, such as P-wave duration, are not recovered by any model. Promising next steps include validating cross-institutional transfer and extending the LA framework to further modalities such as laboratory time series and clinical notes. By learning representations that reflect the multimodal, structured nature of clinical reasoning, LAMAE offers a step toward foundation models that support the holistic, integrative patient assessment clinicians perform in practice.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based assistants) for assistance with writing the manuscript text and for code editing during the implementation of our experiments and analysis pipeline. We have not used generative AI tools for formulating the core research ideas, designing the experiments, or generating any of the results, data, or figures reported in this paper; these are not applicable to this work. Additionally, we used generative AI tools for minor, recommended-disclosure tasks such as grammar and phrasing suggestions and LaTeX formatting assistance. We have reviewed all AI-assisted work: all LLM-suggested text was rewritten and fact-checked by the authors to ensure it accurately reflects our methodology and findings, and all LLM-assisted code changes were manually reviewed, tested, and verified for correctness by the authors before being used to produce any results in this paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses secondary, de-identified clinical data from the MIMIC-IV database and its modality-specific derivatives MIMIC-IV-CXR, MIMIC-IV-ECG, and MIMIC-IV-Echo (Johnson et al., 2023; 2019; Gow et al., 2023; 2026), obtained through PhysioNet under a credentialed-access data use agreement that governs permitted uses, prohibits re-identification attempts, and requires completion of human-subjects-research training. We did not collect any new patient data and did not conduct any new human-subjects research; the original data collection and de-identification were approved by the relevant institutional review board and are documented by the dataset providers (Johnson et al., 2023). MIMIC-IV is drawn from a single academic medical center in the United States, and its patient population is not representative of the global population in terms of demographics, socioeconomic status, or disease prevalence. As with any model trained on such data, predictions may reflect and potentially amplify biases present in the underlying clinical documentation and care patterns (e.g., across race, sex, age, or insurance status); we did not conduct a systematic fairness or subgroup audit of LAMAE's predictions in this work. We view such an audit, together with validation on external and more diverse cohorts, as a necessary prerequisite before any deployment is considered. More broadly, LAMAE is a research contribution toward representation learning for multimodal clinical data, not a validated diagnostic or decision-support tool. Tasks such as in-hospital mortality prediction, diagnostic coding, and length-of-stay estimation are used here as benchmarks to probe representation quality, and the reported performance should not be interpreted as evidence of clinical readiness. Any real-world use of models built on this line of work would require regulatory clearance, prospective clinical validation, human-in-the-loop deployment to guard against automation bias, and ongoing monitoring for distributional shift and disparate performance across patient subgroups.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release an anonymized code repository containing the full implementation of LAMAE, the data processing pipeline, and the training and evaluation scripts at `https://anonymous.4open.science/r/ICLR2027-LAMAE/`. The code will be publicly released upon acceptance. The model architecture, including the LA module and the entity masking scheme, is described in section 3. The construction of our multimodal MIMIC-IV cohort is described in Section 4.1, with the dataset splitting and statistics on modality availability and label distributions detailed in Section C. Experimental setups, including task definitions, splits, baselines, and evaluation protocols, are detailed in Section A. All datasets used (MIMIC-IV and its modality-specific derivatives) are publicly available to credentialed researchers through PhysioNet, allowing independent verification of our data processing and evaluation pipeline.

REFERENCES

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01981-2. URL `http://dx.doi.org/10.1038/s41591-022-01981-2`.

Andrea Agostini, Sonia Laguna, Alain Ryser, Samuel Ruiperez-Campillo, Moritz Vandenhirtz, Nicolas Deperrois, Farhad Nooralahzadeh, Michael Krauthammer, Thomas M Sutter, and Julia E Vogt. Leveraging the structure of medical data for improved representation learning. *International Conference on Machine Learning Workshop on FM4LS*, 2025.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, and others. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision (ECCV)*, pp. 348–367, 2022. doi: 10.1007/978-3-031-19836-6_20.

Alice Bizeul, Thomas Sutter, Alain Ryser, Bernhard Schölkopf, Julius von Kügelgen, and Julia E. Vogt. From Pixels to Components: Eigenvector Masking for Visual Representation Learning, February 2025. URL http://arxiv.org/abs/2502.06314. arXiv:2502.06314 [cs].

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Robert A Byrne, Xavier Rossello, J J Coughlan, Emanuele Barbato, Colin Berry, Alaide Chieffo, Marc J Claeys, Gheorghe-Andrei Dan, Marc R Dweck, Mary Galbraith, Martine Gilard, Lynne Hinterbuchner, Ewa A Jankowska, Peter Jüni, Takeshi Kimura, Vijay Kunadian, Margret Leosdottir, Roberto Lorusso, Roberto F E Pedretti, Angelos G Rigopoulos, Maria Rubini Gimenez, Holger Thiele, Pascal Vranckx, Sven Wassmann, Nanette Kass Wenger, Borja Ibanez, and ESC Scientific Document Group. 2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of acute coronary syndromes of the European Society of Cardiology (ESC). *European Heart Journal*, 44 (38):3720–3826, October 2023. ISSN 0195-668X. doi: 10.1093/eurheartj/ehad191. URL https://doi.org/10.1093/eurheartj/ehad191.

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. PmLR, 2020.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, et al. Radvlm: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333*, 2025.

Kaoru Dohi. Echocardiographic assessment of cardiac structure and function in chronic renal disease. *Journal of Echocardiography*, 17(3):115–122, September 2019. ISSN 1880-344X. doi: 10.1007/s12574-019-00436-x.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Lucas Erlacher, Andrea Agostini, Samuel Ruiperez-Campillo, Ece Ozkan, Thomas M Sutter, and Julia E Vogt. Swissbeatsnet: A multilead masked autoencoder for chagas disease detection. *Challenge*, 15:16, 2025.

Robert B Fetter, Youngsoo Shin, Jean L Freeman, Richard F Averill, and John D Thompson. Case mix definition by diagnosis-related groups. *Medical care*, 18(2):i–53, 1980.

Yuan Gao, Sangwook Kim, Jianzhong You, and Chris McIntosh. ProbMED: A Probabilistic Framework for Medical Multimodal Binding. pp. 20157–20167, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Gao_ProbMED_A_Probabilistic_Framework_for_Medical_Multimodal_Binding_ICCV_2025_paper.html.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All, May 2023. URL http://arxiv.org/abs/2305.05665. arXiv:2305.05665 [cs.CV].

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Ahram Han, Jonathan W Waks, Alistair Johnson, Elizabeth Herbst, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECHO: Echocardiogram Matched Subset. *PhysioNet*, March 2026. doi: 10.13026/nrjh-5r77. URL https://doi.org/10.13026/nrjh-5r77.

Kaiming He, Haoqi Fan, Yuxin Wu, et al. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

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.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

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, January 2019. URL http://arxiv.org/abs/1901.07031. arXiv:1901.07031 [cs].

Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, and Yi Guo. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, August 2024. ISSN 1361-8415. doi: 10.1016/j.media.2024.103202. URL https://www.sciencedirect.com/science/article/pii/S1361841524001270.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *International conference on learning representations*, volume 2025, pp. 8207–8227, 2025.

Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Anthony H Kashou, Peter A Noseworthy, Thomas J Beckman, Nandan S Anavekar, Michael W Cullen, Kurt B Angstman, Benjamin J Sandefur, Brian P Shapiro, Brandon W Wiley, Andrew M Kates, et al. Ecg interpretation proficiency of healthcare professionals. *Current problems in cardiology*, 48(10):101924, 2023.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

Sekeun Kim, Pengfei Jin, Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. Echofm: Foundation model for generalizable echocardiogram analysis, 2025. URL https://arxiv.org/abs/2410.23413.

Sonia Laguna, Andrea Agostini, Alain Ryser, Samuel Ruiperez-Campillo, Irene Cannistraci, Moritz Vandenhirtz, Stephan Mandt, Nicolas Deperrois, Farhad Nooralahzadeh, Michael Krauthammer, et al. Structure is supervision: Multiview masked autoencoders for radiology. *arXiv preprint arXiv:2511.22294*, 2025.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Michelle Espranita Liman, Özgün Turgut, Alexander Müller, Eimo Martens, Daniel Rueckert, and Philip Müller. Echo2ecg: Enhancing ecg representations with cardiac morphology from multi-view echos. *arXiv preprint arXiv:2603.08505*, 2026. Accepted at MICCAI 2026.

Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, John G F Cleland, Andrew J S Coats, Maria G Crespo-Leiro, Dimitrios Farmakis, Martine Gilard, Stephane Heymans, Arno W Hoes, Tiny Jaarsma, Ewa A Jankowska, Mitja Lainscak, Carolyn S P Lam, Alexander R Lyon, John J V McMurray, Alexandre Mebazaa, Richard Mindham, Claudio Muneretto, Massimo Francesco Piepoli, Susanna Price, Giuseppe M C Rosano, Frank Ruschitzka, Anne Kathrine Skibelund, and ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*, 42(36):3599–3726, September 2021. ISSN 0195-668X. doi: 10.1093/eurheartj/ehab368. URL https://doi.org/10.1093/eurheartj/ehab368.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Masoud Mokhtari, Neda Ahmadi, Teresa S. M. Tsang, Purang Abolmaesumi, and Renjie Liao. GEMTrans: A General, Echocardiography-based, Multi-Level Transformer Framework for Cardiovascular Diagnosis, August 2023. URL http://arxiv.org/abs/2308.13217. arXiv:2308.13217 [cs].

Jonathan B Moody, Alexis Poitrasson-Rivière, Jennifer M Renaud, Tomoe Hagio, Fares Alahdab, Mouaz H Al-Mallah, Michael D Vanderver, Sascha N Goonewardena, Edward P Ficaro, and Venkatesh L Murthy. A foundation transformer model with self-supervised learning for ecg-based assessment of cardiac and coronary function. *NEJM AI*, 2(12):AIoa2500164, 2025.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International conference on learning representations*, 2024.

Nader Nemati. Comparative analysis of data augmentation for clinical ecg classification with star. *medRxiv*, pp. 2025–10, 2025.

Lucas Plagwitz, Florian Doldi, Jannes Magerfleisch, Maxim Zotov, Lucas Bickmann, Dominik Heider, Julian Varghese, Lars Eckardt, and Antonius Büscher. QTcNet: a deep learning model for direct heart rate corrected QT interval estimation. *EP Europace*, 27(12):euaf274, December 2025. ISSN 1532-2092. doi: 10.1093/europace/euaf274. URL https://doi.org/10.1093/europace/euaf274.

Gregory A. Roth, Global Burden of Cardiovascular Diseases, and Risks 2023 Collaborators. Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990-2023. *Journal of the American College of Cardiology*, 86(22):2167–2243, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma Technical Report, April 2026. URL http://arxiv.org/abs/2507.05201. arXiv:2507.05201 [cs.AI].

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Nils Strodthoff, JM Lopez Alcaraz, and W Haverkamp IV. Mimic-iv-ecg-ext-icd: Diagnostic labels for mimic-iv-ecg (version 1.0. 1). *PhysioNet. RRID: SCR_007345 https://doi. org/10.13026/hdyc-1h77*, 2024.

Thomas Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia Vogt, Babak Shahbaba, and Stephan Mandt. Unity by diversity: Improved representation learning for multimodal vaes. *Advances in Neural Information Processing Systems*, 37:74262–74297, 2024.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022. Publisher: Nature Publishing Group UK London.

Takeshi Tohyama, Ahram Han, Dukyong Yoon, Kenneth Paik, Brian Gow, Nura Izath, Jacques Kpodonu, and Leo Anthony Celi. Multi-View Echocardiographic Embedding for Accessible AI Development. *medRxiv*, pp. 2025.08.15.25333725, October 2025. doi: 10.1101/2025.08.15.25333725. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12393585/.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, October 2022. URL http://arxiv.org/abs/2203.12602. arXiv:2203.12602 [cs].

Özgün Turgut, Philip Müller, Paul Hager, Suprosanna Shit, Sophie Starck, Martin J. Menten, Eimo Martens, and Daniel Rueckert. Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging. *Medical Image Analysis*, 101: 103451, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,

2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01398. URL https://ieeexplore.ieee.org/document/10203656/.

World Health Organization. *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. World Health Organization, Geneva, 2nd edition, 2004. URL https://who.int.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600, 2023.

Xuan Yang, Rui Xu, Xinchen Ye, Zhihui Wang, Miao Zhang, Yi Wang, Xin Fan, Hongkai Wang, Qingxiong Yue, Xiangjian He, and Yen-Wei Chen. EchoCardMAE: Video Masked Auto-Encoders Customized for Echocardiography. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 171–180, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-05169-1. doi: 10.1007/978-3-032-05169-1_17.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, April 2026. URL http://arxiv.org/abs/2604.18570. arXiv:2604.18570 [cs.LG].

Ziyang Zhang, Qinxin Wu, Sirui Ding, Xiaolong Wang, and Jiancheng Ye. Echo-Vision-FM: A Pre-training and Fine-tuning Framework for Echocardiogram Videos Vision Foundation Model, October 2024. URL https://www.medrxiv.org/content/10.1101/2024.10.09.24315195v2. Pages: 2024.10.09.24315195.

A Implementation Details

Instantiating the study–view–entity hierarchy. Each sample $\mathcal{S}^{(s)}$ is one MIMIC-IV hospital stay, with modalities $\mathcal{M} = \{\text{CXR}, \text{ECG}, \text{Echo}\}$. Hospital stays contain only a subset of these modalities, as depicted in Figure 4. Table 4 shows how each modality is mapped onto the hierarchy of hospital stays as outlined in Section 3. For **CXR**, each radiograph of the stay is a view (frontal, lateral or other). The image itself is the base entity, so $E_{\text{CXR}} = 1$. We use up to three views per study, sampled at random if more are available. For **ECG**, a study is one 12-lead recording sampled at random from the stay. Its base entities are the 12 leads, each a 10 s signal at 500 Hz. For **Echo**, a study is one ultrasound video sampled at random from the stay's examination. Its base entities are 8 frames, linearly spaced over a random window of up to 32 consecutive frames. Missing modalities and views are padded and ignored in both attention and the loss.

Table 4: **Study–view–entity hierarchy for each modality.** V_k: maximum number of views per study; E_k: base entities per view; T_k: tokens per base entity.

	CXR	ECG	Echo
View $X^{(s,k,v)}$	radiograph	12-lead recording	video
Base entity $\boldsymbol{x}^{(s,k,v,e)}$	image	lead	frame
V_k / E_k	3 / 1	1 / 12	1 / 8
Input per entity	$1 \times 224 \times 224$	1×5000	$3 \times 224 \times 224$
Patch size	16×16	50 samples	14×14
T_k	196	100	256

Architecture. Each modality has its own encoder E_{ϕ_k} and decoder D_{θ_k}. The encoders follow ViT-B (12 blocks, width 768, 12 heads), and their transformer blocks are initialized from DINOv3 (Siméoni et al., 2025). The decoders are lighter, with 4 blocks of the same width. Patches are embedded by a linear layer, and each base entity gets its own CLS token. We add learned embeddings for patch position, entity (ECG lead, echo frame) and view (CXR projection, echo view). The implementation of the latent attention module LA_ϕ is a single layer transformer block that operates on the concatenated patch and CLS tokens of all observed modalities. It additionally carries a learned multimodal CLS token, which is not used for downstream tasks. Learned modality embeddings let it distinguish tokens from different modalities. The Independent MAE uses the same encoders and decoders but no LA module.

Masking and objective. We mask a fraction α_k of the tokens of each view. For LAMAE, part of this masking consists of dropping complete entities. The remaining visible entities are masked more lightly, so that the overall ratio stays at α_k. Table 5 lists α_k and the number of dropped entities for each model. The loss is the mean squared error over masked patches, averaged within each modality and then summed over the observed modalities.

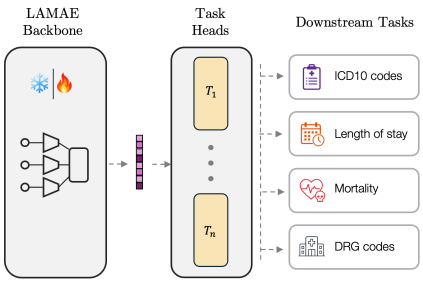


Figure 3: **Finetuning pipeline.** The hospital-stay downstream tasks are shown as an illustrative example. The same finetuning procedure is used for the unimodal downstream tasks.

Downstream heads. For finetuning, we discard the decoders and disable masking. A stay is represented by the average of the entity-level CLS embeddings after the LA module (section 3.4); for the Independent MAE, we average the entity-level CLS embeddings of the encoders. In preliminary experiments, this average performed slightly better than the multimodal CLS token of the LA module, consistent with (He et al., 2022), who find that average pooling works similarly well as a class token. A linear head follows (preceded by LayerNorm). The losses are binary cross-entropy for binary and multi-label tasks, cross-entropy for multi-class tasks and the L1 loss for regression tasks. For unimodal evaluation of LAMAE_{MM}, we feed only the respective modality.

Preprocessing and augmentation. Images are normalized with dataset statistics. CXR training uses random resized crops, horizontal flips and brightness/contrast jitter. Echo training uses random resized crops and brightness jitter, without flips. ECGs are z-score normalized and augmented with STAR time warping (Nemati, 2025).

B Training Details

All models are trained with AdamW (weight decay 0.05), gradient clipping at 1.0 and FP32 precision. The learning rate is scaled with the global batch size B as $\eta = \eta_{\text{base}}\sqrt{B}/256$. Pretraining uses a linear warm-up, a constant learning rate and a short final cooldown. Finetuning uses a linear warm-up followed by cosine decay. Table 5 summarizes the main hyperparameters for pretraining. We use the final pretraining checkpoint. For finetuning, we select the checkpoint with the lowest validation loss and report mean and standard deviation over three seeds. All experiments ran on NVIDIA GH200 GPUs (four per node). Pretraining LAMAE with all the available modalities required approximately *4198* GPU hours, while full finetuning for a single downstream task and seed required approximately *4.33* GPU hours.

Table 5: **Pretraining hyperparameters.** LA: LAMAE; Ind.: Independent MAE. For the hospital-stay comparison (Table 2), $\text{LAMAE}_{\text{CXR}}$ is pretrained for 3000 epochs to give the unimodal CXR baseline its strongest configuration; all models in the unimodal evaluation (Table 3) use 1000-epoch checkpoints for fairness.

	Multimodal		CXR		Echo		ECG	
	LA	Ind.	LA	Ind.	LA	Ind.	LA	Ind.
Epochs	1000	1000	1000	3000/1000	1000	1000	1000	1000
GPUs	128	128	128	128	16	16	128	128
Global batch size	4096	4096	4096	4096	256	256	8192	8192
Base LR	5e-4	5e-4	1e-3	1e-3	4e-3	4e-3	1e-4	1e-4
Mask ratio α_k	0.95	0.90	0.90	0.90	0.875	0.875	0.90	0.90
Dropped ECG leads	8–11	–	–	–	–	–	8–11	–
Dropped Echo frames	6–7	–	–	–	3–4	–	–	–

Table 6: **Finetuning hyperparameters for unimodal downstream tasks.** All settings start from the final pretraining checkpoint, fully finetune the model with a linear warm-up followed by cosine decay, and are repeated with three seeds.

	CXR tasks	CXR ICD-10	Echo LVEF	Echo ICD-10	ECG intervals	ECG ICD-10
Training samples	all	50k	50k	50k	50k	50k
Epochs	50	50	500	40	20	20
GPUs	32	16	32	32	128	128
Global batch size	1024	512	256	256	4096	4096
Base LR	1e-4	1e-4	5e-4	5e-4	1e-4	1e-4
Warm-up steps	500	500	300	300	300	300
Loss	BCE	BCE	L1	BCE	L1	BCE

C Data

Dataset splitting. We split the data at the patient level into training, validation and test sets (80/10/10), so that all stays of a given patient fall into the same split and no patient appears in more than one. Throughout this work, we use the term "hospital stay" to collectively refer to MIMIC emergency department, inpatient, and outpatient encounters. Of the 1,267,017 stays in the MIMIC cohort, 608,170 (48%) contain no ECG, echocardiogram or chest X-ray and are excluded from the analysis. The remaining stays give 138,250 patients (527,805 stays) for training, 17,263 patients (65,296 stays) for validation and 17,321 patients (65,746 stays) for testing. The same split is used for pretraining and all downstream fine-tuning, so no test patient is seen at any stage of training.

Table 7: **Finetuning hyperparameters for hospital-stay downstream tasks.** The same configuration is used for all six tasks: in-hospital mortality, ICD-10 Chapter IX, DRG code, DRG severity, DRG mortality risk, and length of stay. All experiments use 50,000 labeled hospital stays, start from the final pretraining checkpoint, and are repeated over three seeds.

	Hospital-stay tasks
Training samples	50k
Epochs	10
GPUs	32
Global batch size	256
Base LR	5e-4
Warm-up steps	500

Modality availability. Figure 4 summarizes the availability and overlap of CXR, ECG, and Echo across the hospital-stay cohort. Modality availability is highly heterogeneous: hospital stays may contain any subset of the three modalities rather than a complete multimodal examination. This distribution motivates our formulation of each hospital stay as a variable-size set of observed modalities and the explicit handling of missing modalities in LAMAE.

ICD-10 label distributions across modality subsets. Figure 5 shows the distribution of ICD-10 Chapter IX labels across hospital stays with different available modalities. For each modality subset, we report the prevalence and number of positive hospital stays for each retained diagnosis code or code group. Although the absolute number of available samples varies substantially across modality subsets, the plots provide context on the diagnostic composition of the corresponding cohorts and on the class imbalance of the ICD-10 classification task. These distributions complement the per-label ICD-10 classification results reported in Figure 6.

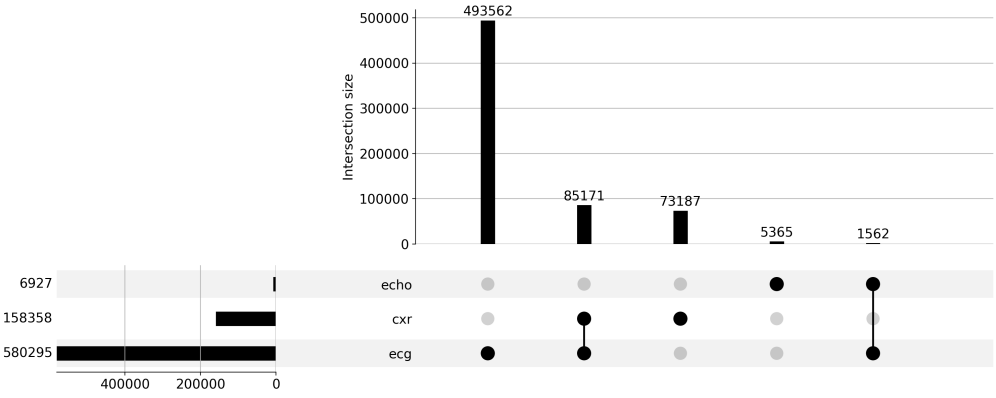
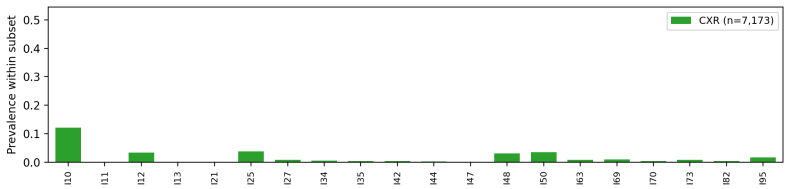


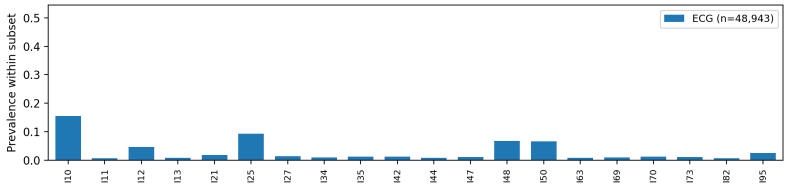
Figure 4: **Modality availability across hospital stays.** UpSet plot showing the availability and overlap of ECG, CXR, and Echo across hospital stays in the multimodal MIMIC-IV cohort. Horizontal bars report the total number of hospital stays containing each modality, while vertical bars report the number of hospital stays for each modality combination.

D Experiments

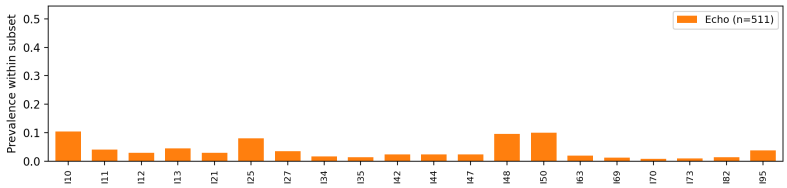
Baselines ProbMED is trained on combinations of CXR, ECG, Echo, and corresponding text samples using a contrastive training objective; it is also trained on data coming from the MIMIC database (Johnson et al., 2023). We note that ProbMED was pretrained on MIMIC using a different data split, such that its pretraining data may overlap with samples in our downstream evaluation split, potentially favoring its performance in this comparison. MedSigLip is the encoder of the MedGemma model family; it natively handles CXR views, as it was trained on de-identified medical data, including CXRs, dermatology images, ophthalmology images, and histopathology slides (Sellergren



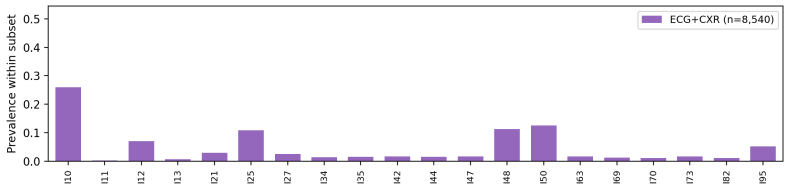
(a) CXR



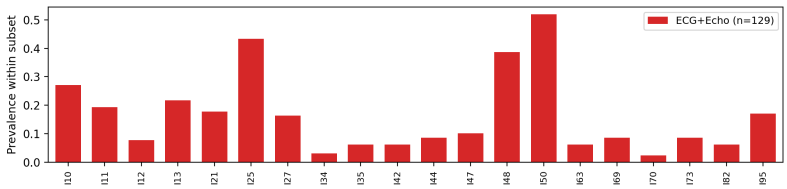
(b) ECG



(c) Echo



(d) ECG + CXR



(e) ECG + Echo

Figure 5: **ICD-10 Chapter IX label distributions across modality subsets.** Prevalence and number of positive hospital stays for each retained ICD-10 Chapter IX diagnosis code or code group, stratified by available modality subset. The distributions illustrate both the differences in sample availability across modality subsets and the class imbalance of the downstream ICD-10 classification task.

et al., 2026). We finetune the MedSigLip encoder with frequency images of ECG signals such that it also has access to all three modalities during finetuning and testing.

Hospital-stay tasks. All hospital-stay tasks are defined per MIMIC-IV admission. *In-hospital mortality* is a binary task that predicts whether the patient dies during the admission. *ICD-10 Chapter IX* (diseases of the circulatory system) is a multi-label task over the diagnosis codes assigned at discharge. We keep all Chapter IX codes and code groups, down to a granularity of five characters, that occur in at least 1% of hospital stays. *DRG* comprises three multi-class tasks derived from the stay: the DRG code itself, the severity of illness and the risk of mortality. *Length of stay* is a three-class task with the classes < 2, 2–7 and > 7 days. We use binary cross-entropy for the binary and

multi-label tasks and cross-entropy for the multi-class tasks. We report the AUROC for in-hospital mortality and the macro-averaged (one-vs-rest) AUROC for all other tasks. The finetuning pipeline for these tasks is illustrated in Fig. 3.

Unimodal tasks. *CXR findings* is a multi-label classification of the 14 CheXpert labels. It is trained with binary cross-entropy, and we report the macro AUROC. *LVEF* is regressed as a fraction in $[0, 1]$ from the structured measurements that accompany the MIMIC-IV-Echo studies. The input is one apical four-chamber video per study, identified with the EchoPrime view classifier. *ECG intervals* jointly regress six global intervals (RR, PR, P-wave duration, QRS duration, QT and QTc) from the 12-lead recording, using one output per interval. We derive the intervals from the machine measurements. Following Plagwitz et al. (2025), we discard records whose annotated intervals fall outside normal physiological bounds (PR < 20 or > 500 ms; P wave < 20 or > 300 ms; QRS < 50 or > 300 ms; QT < 250 or > 600 ms), as well as records carrying the "not measured" sentinel code, since these indicate lower annotation reliability. A record is retained only if all six targets are valid.

Coverage-filled metrics. Because the fine-tuned model uses only a fixed set of input modalities, it cannot make a prediction for evaluation samples that have none of these modalities (see Table 2). Metrics computed only on the covered samples (those with at least one required modality) can therefore look better than the model really is, and are not comparable across modality sets that cover different parts of the cohort. To correct for this, we also report coverage-filled metrics, which are computed on every evaluation sample that has a valid label. For each uncovered sample, the model's output is replaced by a fallback prediction that uses no input data and is estimated from the training split only. For binary classification targets, the fallback is the training prevalence of the positive class, $\hat{p} = \frac{1}{N}\sum_i y_i$ (clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-6}$). It is inserted as the logit $\log \frac{\hat{p}}{1-\hat{p}}$. For multiclass targets, the fallback is the empirical class distribution $\hat{\boldsymbol{\pi}}$ from the training set, inserted as the log-probabilities $\log \hat{\boldsymbol{\pi}}$. For regression targets, it is the mean of the training labels. Formally, the filled prediction for sample i is $\tilde{z}_i = m_i\, z_i + (1 - m_i)\, z^{\text{prior}}$, where $m_i \in \{0, 1\}$ indicates whether any requested modality is available for sample i and z_i is the model output. We then compute the same metrics as in the covered-only evaluation on the filled predictions over the full labelled evaluation set: AUROC, average precision, accuracy and F1 for classification, and MAE and R^2 for regression. A modality set with low coverage is thus penalized in proportion to the fraction of samples it cannot score, because those samples receive an uninformative, constant prediction. As a result, the coverage-filled metrics measure how useful a model is at the population level. As coverage approaches 100%, they converge to the covered-only metrics.

Additional results. To complement the aggregate results reported in the main text, we provide per-label performance for the ICD-10 Chapter IX and CXR classification tasks. Figure 6 reports the test AUROC across individual ICD-10 Chapter IX diagnosis codes and code groups, while Figure 7 reports the test AUROC across the 14 CheXpert labels for $\text{LAMAE}_{\text{CXR}}$, Independent MAE, and LAMAE_{MM}. Figure 8 reports the test R^2 for each of the six ECG interval regression targets (RR, PR, P wave, QRS, QT and QTc) for Independent, $\text{LAMAE}_{\text{ECG}}$ and LAMAE_{MM}, as reported in Table 3c.

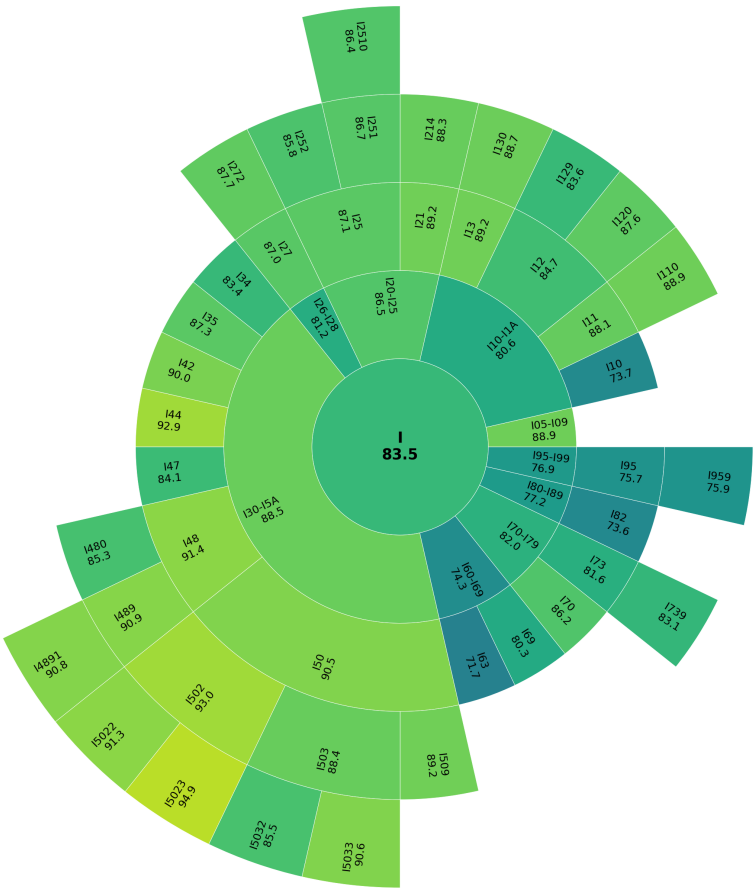


Figure 6: **ICD10 Ch. IX Classification Performance.** Finetuning on the complete MIMIC-IV dataset. Every cell represents an ICD10 code or a group thereof. The center cell is the root, whereas we reach the leaves towards the outer parts of the circle. The leaves are the most fine-grained codes.

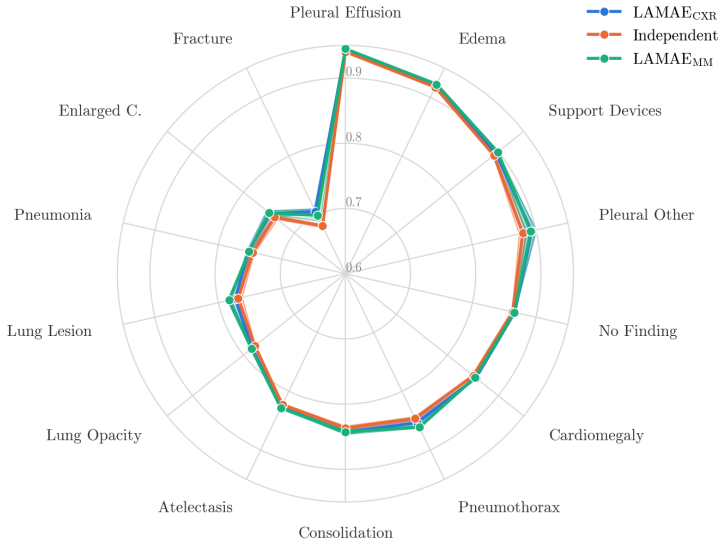


Figure 7: **Test AUROC for each of the 14 chest X-ray (CXR) labels**. Results obtained after finetuning LAMAE$_{CXR}$, Independent and LAMAE$_{MM}$, all three using CXR input only at finetuning. Each point is the mean over three random seeds, and the shaded band shows standard deviation across seeds. Findings are sorted clockwise from the top by mean AUROC across the three models.

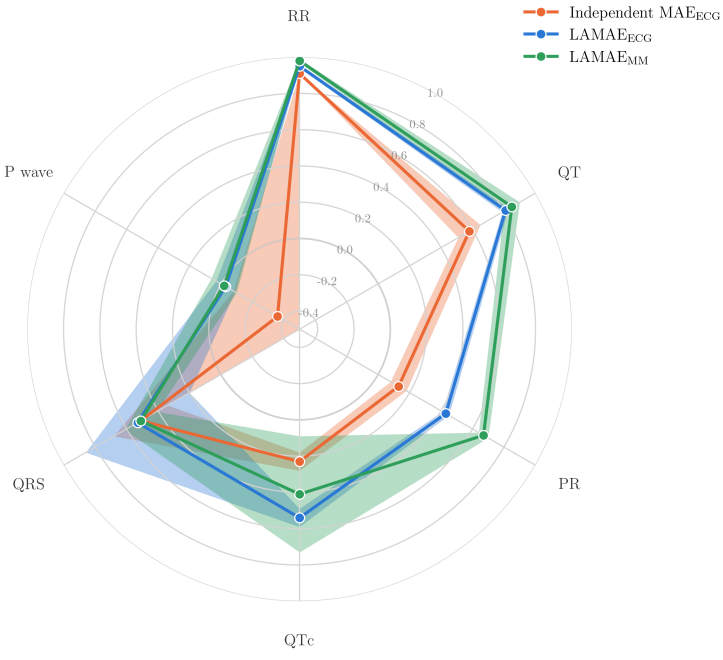


Figure 8: **ECG interval regression.** Test R^2 (↑) per interval (mean ± std) over 3 seeds, Table 3c).