Context-Aware Generative Imputation for Robust Multimodal Learning in Missing Modality Scenarios

Jungwon Choi **Juho Lee**
Kim Jaechul Graduate School of AI, KAIST
{jungwon.choi, juholee}@kaist.ac.kr

Abstract

Multimodal models often suffer from missing data modalities due to acquisition costs, privacy constraints, or sensor failures, leading to severe performance degradation. Existing approaches based on shared representations or expert routing struggle when modality-specific information is absent. When a modality is unobserved, the target representation is not uniquely identifiable, making naive latent prediction prone to degenerate or collapsed solutions. We propose Cross-modal Embedding Prediction and Alignment (CEPA), a framework that addresses missing modalities through task-driven latent representation imputation rather than raw input synthesis. CEPA employs masked representation learning with data-driven masking patterns and a context-conditional distribution alignment objective that stabilizes latent prediction and mitigates representation collapse. We evaluate CEPA on the MIMIC multimodal benchmark spanning EHR time-series, chest X-rays, and radiology reports, across three clinical prediction tasks under controlled missing-modality patterns, and on the CMU-MOSEI sentiment benchmark. CEPA achieves comparable or better performance than prior missing-modality methods, with ablations confirming the contribution of adaptive masking and context-conditional alignment. Code is available at https://github.com/jungwon-choi/CEPA.

1 Introduction

Multimodal learning integrates complementary information from heterogeneous sources such as images, text, and time-series signals [1, 2], but most frameworks assume all modalities are present at both training and inference. In practice, this assumption rarely holds. Modalities are absent due to acquisition costs, privacy restrictions, sensor failures, or data linkage [3–5], and incomplete inputs often cause severe performance degradation [6, 7]. In real deployment, these absences are typically informative rather than uniformly random. A chest X-ray may be skipped when a patient is clinically stable, and a sensor stream may be dropped when its readings appear redundant. These patterns motivate evaluating both controlled missing-modality protocols and the naturally observed modality availability.

Existing approaches to missing-modality scenarios fall into three paradigms. *Generative imputation* [8] reconstructs missing raw data through autoencoders, but is expensive for high-dimensional modalities and exhibits semantic drift that degrades downstream tasks. *Feature alignment* [9, 4] projects modalities into a shared space, but fails when modalities carry disjoint information or different temporal resolutions. *Domain-specific architectures* [3–5] embed task-specific inductive biases that limit applicability to new domains. Other strategies, including meta-learning [6], disentanglement [9, 5, 10], and contrastive learning [11], extend these paradigms but inherit limitations.

Across these paradigms, missing-modality learning is cast as a reconstruction or corruption problem. When a modality is unobserved, however, there is no uniquely identifiable ground-truth representation, so pointwise reconstruction objectives are ill-posed and often induce representation collapse. Distribution-level supervision is therefore a more natural target than pointwise reconstruction.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we introduce CEPA, **C**ross-modal **E**mbedding **P**rediction and **A**lignment, a framework that imputes missing-modality embeddings in latent space rather than reconstructing raw inputs. Inspired by masked representation learning [12, 13], CEPA predicts missing-modality embeddings from observed ones. We propose an *adaptive cross-modal masking strategy* that learns conditional masking patterns, and a *context-conditional distribution alignment* objective that aligns predicted embeddings with the distribution of observed ones for contextually similar samples, acting as a task-coupled regularizer that mitigates representation collapse.

Our main contributions are as follows (related work is in Appendix C):

- We propose CEPA, which learns adaptive cross-modal masking patterns and uses them to train a cross-modal embedding predictor that imputes missing modalities at the latent level, avoiding the cost and semantic drift of raw-input reconstruction.
- We introduce a context-conditional characteristic-function alignment loss that regularizes predicted embeddings against context-similar observations, mitigating the representation collapse common to ground-truth-free imputation.
- On three MIMIC clinical tasks and the CMU-MOSEI sentiment benchmark, CEPA achieves comparable or better performance than recent missing-modality baselines under controlled evaluation.

2 Methods

To handle missing modalities, CEPA predicts their representations from the observed ones instead of reconstructing raw inputs (Figure 1). CEPA combines three components: an *adaptive masking* module, a *representation predictor* g_ϕ, and a *context-conditional characteristic-function (CF) alignment* term. First, the masking module hides a subset of the informative observed tokens, which discourages trivial solutions and pushes the predictor to exploit cross-modal correlations. Second, the predictor fills in representations for both these hidden tokens and the fully missing modalities through cross-modal interactions. Third, the CF term keeps these imputations distributionally consistent with the observed embeddings of contextually similar samples, regularizing the predictor against collapse.

2.1 Problem Setup and Notations

Consider a multimodal dataset consisting of N data points, where each point may have only a subset of the M modalities observed. Let $[M] := \{1, \ldots, M\}$ denote the set of all modalities and $[N] := \{1, \ldots, N\}$ the set of data indices. For each $i \in [N]$, we denote by $\mathcal{O}_i \subseteq [M]$ the set of modalities observed for the i^{th} data point, and the set of missing modalities as $\mathcal{M}_i := [M] \setminus \mathcal{O}_i$. The dataset is then given by $\mathcal{D} = \{(\mathbf{X}_i^{(\mathcal{O}_i)}, y_i)\}_{i=1}^N$ where $\mathbf{X}_i^{(\mathcal{O}_i)} := \{\mathbf{X}_i^{(m)}\}_{m \in \mathcal{O}_i}$ collects the observed modality-specific inputs for the i^{th} point, $\mathbf{X}_i^{(m)}$ denotes the observation from the m^{th} modality corresponding to the i^{th} input, and y_i denotes the label corresponding to $\mathbf{X}_i$. For notational simplicity, we omit the sample index i when the context is clear.

We consider a collection of modality-specific encoders $\{\mathcal{E}^{(m)}\}_{m \in [M]}$, where $\mathcal{E}^{(m)}$ denotes the encoder associated with the m^{th} modality. We assume that each modality-specific tokenized input $\mathbf{X}^{(m)}$ admits a natural sequential structure, represented as $\mathbf{X}^{(m)} := [\mathbf{x}_1^{(m)}, \ldots, \mathbf{x}_{L^{(m)}}^{(m)}]$ where $\mathbf{x}_t^{(m)}$ denotes the t^{th} token and $L^{(m)}$ is the sequence length. Such sequential structures arise naturally across diverse data modalities, including image patches in vision models, (subword) tokens in text, and temporal segments in time-series data. Each modality is processed by its corresponding encoder $\mathcal{E}^{(m)}$, yielding a representation $\mathbf{Z}^{(m)} = [\mathbf{z}_1^{(m)}, \ldots, \mathbf{z}_{L^{(m)}}^{(m)}] \in \mathbb{R}^{L^{(m)} \times d_m}$ where d_m denotes the dimensionality of the token-level features for modality m.

When some modalities are missing, multimodal models must handle incomplete inputs while still producing reliable predictions. A common strategy is to impute missing values with trivial substitutes (e.g., zeros or learnable parameters), but this approach fails to capture important cross-modal relationships that influence the final prediction. To address this, we propose learning a *representation predictor* g_ϕ with parameters ϕ, which takes encoded representations from observed modalities $\{\mathbf{Z}^{(m)}\}_{m \in \mathcal{O}}$ and predicts the missing representations $\{\mathbf{Z}^{(m)}\}_{m \in \mathcal{M}}$ leveraging cross-modal interactions. To achieve this, we introduce CEPA, a simple yet effective cross-modal learning framework for training the representation predictor.

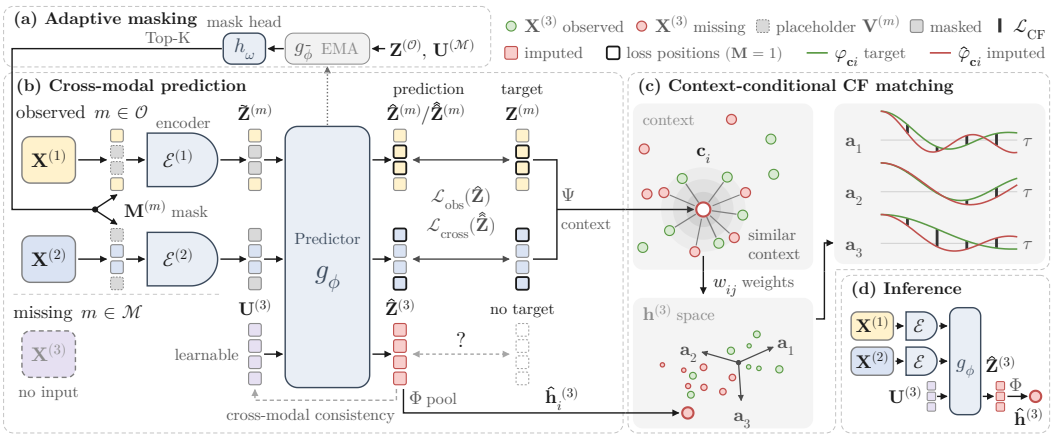


Figure 1: **Overview of CEPA** (modalities 1 and 2 observed, modality 3 missing). **(a) Adaptive masking.** The mask head h_ω, guided by the EMA predictor $g_{\bar{\phi}}$, masks the Top-K tokens of each observed modality (Figure 2). **(b) Cross-modal prediction.** g_ϕ predicts the embeddings of all modalities from the masked embeddings $\tilde{\mathbf{Z}}^{(m)}$ and the learnable tokens $\mathbf{U}^{(3)}$, trained with $\mathcal{L}_{\text{obs}}$ and the cross-modal consistency loss $\mathcal{L}_{\text{cross}}$. The missing modality has no target. **(c) Context-conditional CF matching.** $\mathcal{L}_{\text{CF}}$ matches the distribution of the imputations to that of the observed embeddings of samples with a similar context $\mathbf{c}_i$ (Figure 8). **(d) Inference.** g_ϕ imputes $\widehat{\mathbf{Z}}^{(3)}$ from unmasked inputs.

2.2 Predicting Missing Representations

To learn the predictor g_ϕ, we adopt a masked representation learning approach that learns the model by reconstructing the representations from masked inputs [12]. Instead of predicting raw signals, we learn the predictor to reconstruct in representation space, thereby avoiding wasted capacity on reconstructing minor artifacts or low-level details in high-dimensional spaces and encouraging it to focus on capturing semantic content [13–15].

The overall training pipeline goes as follows. Given an input $\mathbf{X}^{(\mathcal{O})}$, we first apply a mask $\mathbf{M}^{(m)}$ to $\mathbf{X}^{(m)}$ for each $m \in \mathcal{O}$. For instance, if the m^{th} modality were images, the mask may be a binary matrix that masks out some pixels or patches of the images. Then we put the masked input to the encoder $\mathcal{E}^{(m)}$ to get,

$$\tilde{\mathbf{Z}}^{(m)} := [\tilde{\mathbf{z}}_1^{(m)}, \ldots, \tilde{\mathbf{z}}_{L^{(m)}}^{(m)}] = \mathcal{E}^{(m)}\left(\mathbf{X}^{(m)} \odot (1 - \mathbf{M}^{(m)}) + \mathbf{V}^{(m)} \odot \mathbf{M}^{(m)}\right), \tag{1}$$

where $\mathbf{M}^{(m)} \in \{0,1\}^{L^{(m)}}$ is a binary mask with $\mathbf{m}_t^{(m)} = 1$ indicating that the t^{th} token is masked (i.e., replaced by the learnable placeholder $\mathbf{V}^{(m)}$) and $\mathbf{m}_t^{(m)} = 0$ a visible token. For a missing modality $m \in \mathcal{M}$, we start from a learnable token $\mathbf{u}^{(m)} \in \mathbb{R}^{d_m}$ that is replicated to form $\mathbf{U}^{(m)} = [\mathbf{u}^{(m)}, \ldots \mathbf{u}^{(m)}]^\top \in \mathbb{R}^{L^{(m)} \times d_m}$.

Finally, before putting the representations into the predictor g_ϕ, for each modality m, we augment the representation with a learnable modality embedding $\mathbf{e}^{(m)} \in \mathbb{R}^{d_m}$ as replicated form of $\mathbf{E}^{(m)} = [\mathbf{e}^{(m)}, \ldots \mathbf{e}^{(m)}]^\top \in \mathbb{R}^{L^{(m)} \times d_m}$ to distinguish between different modalities. We also add sinusoidal positional embeddings $\mathbf{P}^{(m)} \in \mathbb{R}^{L^{(m)} \times d_m}$ to encode position information within each modality. For notational simplicity, we omit these embeddings in subsequent equations, but they are always applied before concatenation. During the training, the predictor g_ϕ takes these inputs and predicts the representations from the missing modalities as well as the representations from the unmasked inputs in the observed modalities,

$$(\widehat{\mathbf{Z}}^{(m)})_{m \in [M]} := g_\phi\Big(\texttt{concat}\big(\{\tilde{\mathbf{Z}}^{(m)}\}_{m \in \mathcal{O}}, \{\mathbf{U}^{(m)}\}_{m \in \mathcal{M}}\big)\Big). \tag{2}$$

In a typical masked representation learning framework, masks are sampled at random from a simple distribution (e.g., a Bernoulli distribution for binary masks), and the model is trained to minimize the discrepancy between $\widehat{\mathbf{Z}}^{(m)}$ and the representation $\mathbf{Z}^{(m)}$ computed from the unmasked inputs. In our setting, however, two key challenges arise:

- **Random masks ignore cross-modal correlations:** In multimodal data, different modalities share common semantic content, giving rise to cross-modal correlations. Rather than masking inputs at random as in single-modality settings, we can exploit these correlations—for example, by designing masks for a specific modality guided by the other modalities.
- **Absence of targets for missing modalities:** Unlike the standard case, we lack ground-truth representations $(\mathbf{Z}^{(m)})_{m\in\mathcal{M}}$ for missing modalities. Simply skipping learning for these modalities may be ineffective, particularly at high missing ratios.

In the following sections, we describe our method to tackle these challenges.

2.3 Cross-modal Embedding Prediction Architecture

2.3.1 Masking the observed modalities

A straightforward approach is to train the representation predictor g_ϕ to impute missing-modality representations directly from the observed ones. However, for truly missing modalities, no uniquely identifiable ground-truth representations exist, making pointwise reconstruction ill-posed. Moreover, representations that are optimal for missing-modality prediction may not coincide with those that support downstream task performance, potentially leading to representation collapse or task-irrelevant features [16].

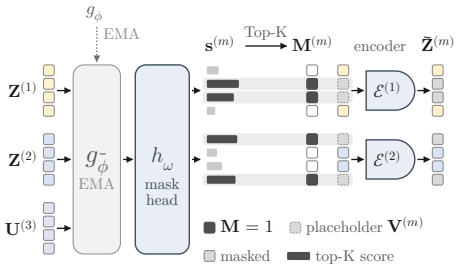


Figure 2: **Adaptive cross-modal masking.** The EMA predictor $g_{\bar{\phi}}$ and the mask head h_ω score every observed token ($\mathbf{s}^{(m)}$), and the Top-$K^{(m)}$ positions form $\mathbf{M}^{(m)}$ and are replaced by $\mathbf{V}^{(m)}$ before the encoder, giving $\tilde{\mathbf{Z}}^{(m)}$.

To obtain a stronger and more stable training signal, we mask inputs from the observed modalities and train g_ϕ to predict representations for both observed and missing modalities. This cross-modal embedding prediction strategy discourages trivial solutions, encourages semantically meaningful cross-modal relationships, and provides auxiliary supervision beyond missing-modality prediction.

Nevertheless, naive random masking can be suboptimal because modalities differ in information density and semantic relevance. We therefore learn which portions of the observed modalities should be masked to best support representation prediction for the missing modalities.

2.3.2 Modality-aware mask prediction

To determine which parts of the observed modalities to mask, we introduce a *mask predictor* h_ω with parameters ω. The mask predictor is guided by an exponential moving average (EMA) of the representation predictor parameters ϕ, which serves as a slowly varying reference, $\bar{\phi} \leftarrow \lambda\bar{\phi} + (1-\lambda)\mathsf{stopgrad}(\phi)$, where $\lambda \in [0,1)$ is the decay coefficient and $\mathsf{stopgrad}(\cdot)$ is the stop gradient operation. Using $\bar{\phi}$, for each observed modality $m \in \mathcal{O}$, we compute

$$\begin{aligned}
\mathbf{s}^{(m)} &= h_\omega\Big(\big[g_{\bar{\phi}}\big(\mathtt{concat}\big(\{\mathbf{Z}^{(m')}\}_{m'\in\mathcal{O}}, \{\mathbf{U}^{(m')}\}_{m'\in\mathcal{M}}\big)\big)\big]^{(m)}\Big),\\
\mathcal{I}^{(m)} &= \mathtt{TopK}(\mathbf{s}^{(m)}, K^{(m)}) \subseteq \{1,\dots,L^{(m)}\},\\
\mathbf{m}_t^{(m)} &= \mathbf{1}[t \in \mathcal{I}^{(m)}],\ \mathbf{M}^{(m)} = [\mathbf{m}_1^{(m)},\dots,\mathbf{m}_{L^{(m)}}^{(m)}],
\end{aligned} \tag{3}$$

where $[\cdot]^{(m)}$ denotes the output tokens of modality m, $\mathtt{TopK}(\cdot, K^{(m)})$ retrieves the token indices with top $K^{(m)}$ activations, $\mathbf{1}[\cdot]$ denotes an indicator function returning one for selected tokens and zero otherwise, and $\mathbf{M}^{(m)} := [\mathbf{m}_1^{(m)},\dots,\mathbf{m}_{L^{(m)}}^{(m)}]$ is the resulting binary mask. Since the $\mathtt{TopK}$ selection and indicator function are non-differentiable, we employ a straight-through estimator (STE) [17] to enable gradient flow: the forward pass uses the hard binary mask, while the backward pass propagates gradients through the continuous activation scores $\mathbf{s}^{(m)}$ produced by h_ω. With the adaptively determined mask $\mathbf{M}^{(m)}$, masked tokens are replaced with learnable placeholders $\mathbf{V}^{(m)}$, and the resulting input is encoded into $\tilde{\mathbf{Z}}^{(m)}$ (cf. Section 2.2).

The key intuition is that the mask predictor estimates which parts of the representations are most important according to the current state of the model. Conceptually, it maps representations into activation signals, where larger values correspond to more informative components. By selectively masking out these important parts and requiring the model to reconstruct them (and simultaneously predict the missing modalities), the representation predictor g_ϕ is encouraged to capture essential cross-modal relationships and thus learn more robust representations.

2.3.3 Multi-objective representation prediction

Given the representations of the observed modalities $\{\tilde{\mathbf{Z}}^{(m)}\}_{m\in\mathcal{O}}$ computed via Equation (1), the representation predictor g_ϕ predicts the representations for both observed modalities with unmasked inputs and missing modalities, as specified in Equation (2). To train g_ϕ, we adopt distinct objective functions for observed and missing modalities, which we describe in detail below.

Prediction for observed modalities. Training for the observed modalities is straightforward, as the ground-truth representations $\mathbf{Z}^{(m)}$ for $m \in \mathcal{O}$ can be computed directly from unmasked data. Although this objective may appear redundant, given that modality-specific encoders are already available, we find it beneficial in practice, as it encourages the representation predictor to more effectively exploit cross-modal correlations. The objective for the observed modalities is given by

$$\mathcal{L}_{\text{obs}} = \frac{1}{|\mathcal{O}|}\sum_{m\in\mathcal{O}}\frac{1}{K^{(m)}}\left\|\left(\widehat{\mathbf{Z}}^{(m)} - \mathbf{Z}^{(m)}\right)\odot\mathbf{M}^{(m)}\right\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Prediction for missing modalities through cross-modal consistency. Since ground-truth representations are unavailable for missing modalities, we enforce a *cross-modal consistency* objective: the predicted missing representations should contain sufficient semantic information to support reconstruction of the observed ones. Concretely, we compute

$$(\widehat{\widehat{\mathbf{Z}}}^{(m)})_{m\in[M]} := g_\phi\Big(\texttt{concat}\big(\{\tilde{\mathbf{Z}}'^{(m)}\}_{m\in\mathcal{O}}, \{\widehat{\mathbf{Z}}^{(m)}\}_{m\in\mathcal{M}}\big)\Big) \tag{5}$$

that is, the representations predicted with the missing part replaced by the predicted missing representations $\widehat{\mathbf{Z}}^{(m)}$. Also, for the representation $\tilde{\mathbf{Z}}'^{(m)}$, we mask a larger fraction of tokens, $K'^{(m)}/L^{(m)}$, than for $\mathcal{L}_{\text{obs}}$, encouraging g_ϕ to rely more on the predicted missing representations. We then compare the reconstructed outputs $(\widehat{\widehat{\mathbf{Z}}}^{(m)})_{m\in\mathcal{O}}$ against the ground-truth from the unmasked data:

$$\mathcal{L}_{\text{cross}} = \frac{1}{|\mathcal{O}|}\sum_{m\in\mathcal{O}}\frac{1}{K'^{(m)}}\left\|\left(\widehat{\widehat{\mathbf{Z}}}^{(m)} - \mathbf{Z}^{(m)}\right)\odot\mathbf{M}'^{(m)}\right\|_F^2. \tag{6}$$

By minimizing this loss, the model is driven to produce missing representations that are semantically consistent with the observed ones, enabling reliable cross-modal reconstruction.

2.4 Context-Aware Conditional Distribution Alignment

While the cross-modal consistency objective ($\mathcal{L}_{\text{cross}}$) enforces semantic alignment across modalities, it does not explicitly constrain the distributional geometry of the representations. For example, in Equation (5), the representation predictor may rely only weakly on the reconstructed missing representations $\widehat{\mathbf{Z}}^{(m)}$, instead depending primarily on the partially observed inputs $\tilde{\mathbf{Z}}'^{(m)}$. Although we prevent this behavior by increasing the masking ratio for $\tilde{\mathbf{Z}}'^{(m)}$, the predictor may still degenerate, leading to representation collapse. To address this, we introduce a framework that aligns predictions to a dynamic *empirical conditional distribution* derived from contextually similar samples.

Feature pooling and context aggregation. Since different modalities possess varying sequence lengths and feature dimensions, comparing them directly is infeasible. We thus map them into a shared semantic space via modality-specific transformations. For all $m \in [M]$, we define a *modality-specific pooling* $\Phi^{(m)} : \mathbb{R}^{L^{(m)}\times d_m} \to \mathbb{R}^d$ that transforms the variable-length sequence $\mathbf{Z}^{(m)}$ into a fixed-dimensional *pooled representation* $\mathbf{h}^{(m)}$:

$$\mathbf{h}^{(m)} = \Phi^{(m)}(\mathbf{Z}^{(m)}), \quad \widehat{\mathbf{h}}^{(m)} = \Phi^{(m)}(\widehat{\mathbf{Z}}^{(m)}). \tag{7}$$

To aggregate the variable-sized set of observed pooled representations into a fixed-dimensional *context vector* $\mathbf{c} \in \mathbb{R}^d$, we employ a permutation-invariant set encoder Ψ. Specifically, we prepend a learnable query token $\mathbf{q} \in \mathbb{R}^d$ to the set $\{\mathbf{h}^{(m)}\}_{m\in\mathcal{O}}$ and process the resulting sequence through a Transformer encoder:

$$\mathbf{c} = \Psi\big(\mathbf{q},\, \{\mathbf{h}^{(m)}\}_{m\in\mathcal{O}}\big), \tag{8}$$

where Ψ is a permutation-invariant set encoder that summarizes the available modalities via attention with a learnable query. This attention-based aggregation adaptively weights each modality's contribution depending on the available modality configuration, unlike simple mean pooling which treats all observed modalities equally regardless of the observation pattern.

Context-aware empirical target. We assume that the encoders together with the pooling operators induce (for fixed model parameters) a coherent joint distribution $P(\mathbf{c}, \mathbf{h}^{(m)}) = P(\mathbf{h}^{(m)} \mid \mathbf{c})P(\mathbf{c})$, and we treat the resulting pairs $\{(\mathbf{h}_i^{(m)}, \mathbf{c}_i)\}_{i=1}^N$ computed from the dataset as samples from this distribution. Fix a modality m and a minibatch $\mathcal{B} \subseteq [N]$, and let $B_m \subseteq \mathcal{B}$ denote the samples in $\mathcal{B}$ for which modality m is observed. For a query sample i missing modality m, we use kernel regression in the context space to estimate conditional expectations of the form $\mathbb{E}[f(\mathbf{h}^{(m)}) \mid \mathbf{c} = \mathbf{c}_i]$. Concretely, for any test function $f : \mathbb{R}^d \to \mathbb{C}$, we define the Nadaraya–Watson (NW) estimator

$$\mathbb{E}_{B_m}^{\mathrm{NW}}\big[f(\mathbf{h}^{(m)}) \mid \mathbf{c}_i\big] := \sum_{j\in B_m} \underbrace{\frac{\kappa(\mathbf{c}_i, \mathbf{c}_j)}{\sum_{j'\in B_m} \kappa(\mathbf{c}_i, \mathbf{c}_{j'})}}_{:=w_{ij}^{(m)}} f(\mathbf{h}_j^{(m)}), \tag{9}$$

where $\kappa(\mathbf{c}_i, \mathbf{c}_j) = \exp(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{2\sigma^2})$. To prevent degenerate solutions that collapse the context geometry, we detach the weights $w_{ij}^{(m)}$ from the computation graph, i.e., $w_{ij}^{(m)} \leftarrow \mathsf{stopgrad}(w_{ij}^{(m)})$.

Conditional characteristic-function target. We instantiate Equation (9) with the complex exponential test function $\exp(\mathrm{i}\tau\, \mathbf{a}^\top \mathbf{h})$ for a frequency $\tau \in \mathbb{R}$ and a projection vector $\mathbf{a} \in \mathbb{R}^d$ so that $\mathbb{E}[\exp(\mathrm{i}\tau\, \mathbf{a}^\top \mathbf{h}^{(m)}) \mid \mathbf{c}]$ is the conditional characteristic function (CF) of $P(\mathbf{h}^{(m)} \mid \mathbf{c})$. This yields the NW target CF for query context $\mathbf{c}_i$:

$$\varphi_{\mathbf{c}_i}^{(m)}(\tau, \mathbf{a}) := \sum_{j\in B_m} w_{ij}^{(m)}\, \exp\big(\mathrm{i}\tau\, \mathbf{a}^\top \mathbf{h}_j^{(m)}\big). \tag{10}$$

Cross-modal CF matching. Let $B_m^c = \mathcal{B} \setminus B_m$ denote the set of samples for which modality m is missing (including the query sample i), and let $\widehat{\mathbf{h}}_j^{(m)}$ be the pooled representation predicted for modality m on sample $j \in B_m^c$. Using these predicted features, we analogously form an NW estimate of the conditional CF:

$$\widehat{\varphi}_{\mathbf{c}_i}^{(m)}(\tau, \mathbf{a}) := \sum_{i'\in B_m^c} \widehat{w}_{ii'}^{(m)} \exp\big(\mathrm{i}\tau \mathbf{a}^\top \widehat{\mathbf{h}}_{i'}^{(m)}\big), \tag{11}$$

where $\widehat{w}_{ii'}^{(m)}$ is defined as in Equation (9), but normalized over the index set B_m^c.

We regularize the predicted representations by matching the observed-side estimator $\varphi_{\mathbf{c}_i}^{(m)}$ to the predicted-side estimator $\widehat{\varphi}_{\mathbf{c}_i}^{(m)}$, which encourages the conditional distribution of $\widehat{\mathbf{h}}^{(m)}$ given $\mathbf{c}$ to align with that of $\mathbf{h}^{(m)}$ and mitigates representation collapse. Specifically, we define the following CF matching regularizer, which can be viewed as a cross-modal analogue of the SIGReg objective based on Epps–Pulley test functions [18]:

$$\mathcal{L}_{\mathrm{CF}}^{(m)} = \frac{1}{|B_m^c|} \sum_{i\in B_m^c} \sum_{\ell,k=1}^{N_a, N_\tau} \frac{\big|\varphi_{\mathbf{c}_i}^{(m)}(\tau_k, \mathbf{a}_\ell) - \widehat{\varphi}_{\mathbf{c}_i}^{(m)}(\tau_k, \mathbf{a}_\ell)\big|^2}{N_a N_\tau}. \tag{12}$$

Here we average discrepancies over N_a random projections $\{\mathbf{a}_\ell\}_{\ell=1}^{N_a}$ and N_τ sampled frequencies $\{\tau_k\}_{k=1}^{N_\tau}$, as well as over the query samples $i \in B_m^c$ missing modality m. Figure 8 in Appendix F.3 illustrates both estimators on a minibatch. The total CF regularization over all missing modalities is then given by $\mathcal{L}_{\mathrm{CF}} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \mathcal{L}_{\mathrm{CF}}^{(m)}$.

We treat $\mathcal{L}_{\text{CF}}$ as a task-coupled structural regularizer rather than as recovery of an external conditional law. Because the context, kernel geometry, and predictor are learned jointly, the NW estimator does not in general yield a consistent estimate of $P(\mathbf{h}^{(m)} \mid \mathbf{c})$. Its role is to encourage examples with similar observed contexts to receive similarly distributed predicted representations, which mitigates representation collapse and improves downstream robustness.

Practical estimator details. Consistent with the definition of B_m^c above, the query index i is included in the predicted-side estimator $\widehat{\varphi}_{\mathbf{c}_i}^{(m)}$. We skip the CF term for any minibatch in which $|B_m| = 0$. Since the context vectors are ℓ_2-normalized, we use a fixed bandwidth $\sigma = 1.0$ for all main results and report an adaptive per-minibatch median rule in Appendix D.9. With our minibatch sizes, $|B_m|$ is typically several samples, which suffices for a low-frequency CF target. As highlighted by Balestriero and LeCun [18], CF-based matching can offer improved geometric stability compared to density-based divergences, and in the following proposition we show that the gradient of Equation (12) is naturally bounded.

Proposition 2.1 (Gradient Boundedness)**.** *Let $\|\mathbf{a}_\ell\|_2 = 1$ be unit projection directions. The gradient of $\mathcal{L}_{CF}^{(m)}$ with respect to the prediction $\widehat{\mathbf{h}}$ is bounded in terms of the sampled frequencies. Specifically, $\|\nabla_{\widehat{\mathbf{h}}} \mathcal{L}_{CF}^{(m)}\|_2 \leq 4 \max_k |\tau_k|$.*

This bound arises because the characteristic function maps data to the unit circle in the complex plane, inherently limiting the gradient magnitude regardless of the density steepness. The bound is independent of the context distribution and the kernel bandwidth (see Appendix A for the proof).

2.5 Multimodal Fusion and Task Prediction

After obtaining the complete set of multimodal representations—comprising observed representations $\mathbf{h}_i^{(m)}$ for $m \in \mathcal{O}$ and predicted representations $\widehat{\mathbf{h}}_i^{(m)}$ for $m \in \mathcal{M}$—we aggregate them for the final task prediction. Since the pooled representations $\mathbf{h}$ provide a fixed-dimensional summary of each modality, they can be directly concatenated:

$$\mathbf{h}_i^{\text{fused}} = \texttt{concat}\left(\{\mathbf{h}_i^{(m)}\}_{m\in\mathcal{O}}, \{\widehat{\mathbf{h}}_i^{(m)}\}_{m\in\mathcal{M}}\right). \tag{13}$$

The fusion module f_ψ then maps this fused vector to the final prediction $\hat{y}_i = f_\psi(\mathbf{h}_i^{\text{fused}})$. The task loss $\mathcal{L}_{\text{task}}$ is cross-entropy for classification and mean squared error for regression.

Overall objective. The proposed context-conditional distribution alignment term is applied to missing modalities and added to the original training objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{obs}} + \beta\mathcal{L}_{\text{cross}} + \gamma\mathcal{L}_{\text{CF}}, \tag{14}$$

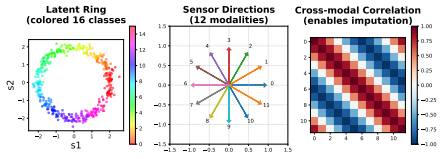
where α, β, and γ control the strength of each loss term. This objective encourages predicted missing representations to match the *context-conditional* distribution of representations obtained from actual observations, thereby mitigating representation collapse while preserving contextual diversity.

3 Experiments

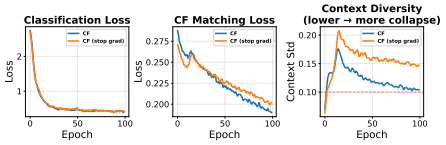
3.1 Synthetic Multimodal Experiments

To isolate the behavior of missing-modality prediction strategies, we design a controlled synthetic experiment with a ring-structured latent space and 12 projected modalities (Figure 3). The latent angle determines the class label, and we report accuracy against the number of observed modalities.

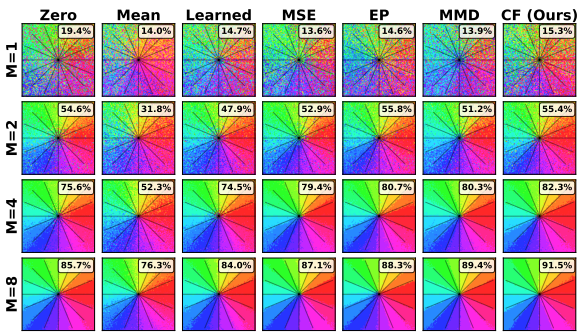
Figure 3(c) compares Zero, Mean, Learned token, supervised MSE, marginal EP, marginal MMD, and our context-conditional CF matching. CF produces the most coherent decision boundaries and achieves the highest average accuracy, especially at moderate numbers of observed modalities. Static fill strategies yield fragmented boundaries, while MSE and marginal MMD underperform because they lack context-conditioned distributional consistency. Figure 3(b) further shows that CF matching without stop-gradient collapses the context representation, whereas our stop-gradient formulation preserves context diversity. Full details are in Appendix E.4.



(a) Data generation setup.



(b) Context collapse analysis.



(c) Decision boundaries across strategies and observed M modalities.

Figure 3: **Synthetic experiment. (a)** Ring-structured latent data with 12 projected modalities. **(b)** Stop-gradient prevents context collapse in CF matching. **(c)** Decision boundaries across fill strategies with $M \in \{1, 2, 4, 8\}$ observed modalities. CF (Ours) yields the most coherent boundaries.

3.2 Datasets and Experimental Setup

Multimodal MIMIC Dataset and Tasks. We utilize the MIMIC-IV and MIMIC-CXR databases [19, 20] with three complementary modalities: structured time-series EHR, chest X-ray images (CXR), and radiology reports (TXT). Our cohort follows the MedFuse protocol [3], built on the Harutyunyan clinical time-series benchmark [21]. EHR measurements are hourly discretized with carry-forward imputation and observation masks. CXR studies are linked to ICU stays at the patient level, and radiology reports are paired one-to-one with CXR studies. Discharge summaries and nursing notes are not used. The fully paired subset is used for controlled missing-modality experiments. The partially paired cohort, in which roughly 74% of ICU stays do not have a paired CXR, is used for evaluation under naturally observed modality absence. We evaluate on three clinical prediction tasks: (1) *in-hospital mortality* (binary classification), (2) *length-of-stay* (regression), and (3) *phenotyping* (multi-label classification of 25 conditions). For mortality and length-of-stay, inputs are restricted to the first 48 hours after ICU admission. Dataset statistics and detailed configurations are provided in Appendix E. Beyond the clinical setting, we additionally evaluate on CMU-MOSEI, a multimodal sentiment benchmark with audio, language, and vision modalities (Section 3.6).

3.3 Baselines

We compare CEPA against 11 missing-modality baselines spanning meta-learning, contrastive learning, disentanglement, imputation, mixture-of-experts, partial-information decomposition, and multimodal VAE approaches: SMIL [6], MUSE [11], ShaSpec [9], DRIM [10], M3Care [4], DiCMoR [22], FuseMoE [23], Flex-MoE [24], SimMLM [7], Robult [25], and MMVM-VAE [26]. We also adapt M3-JEPA [27] as a cross-modal representation learning baseline. Methods requiring discrete labels are excluded from length-of-stay prediction. Adaptation details are in Appendix E.3.

3.4 Implementation Details

Model Architecture. For MIMIC-IV, we employ domain-specific encoders: a Transformer encoder for structured EHR time-series data, SigLIP2 [28] for chest X-ray images, and CXR-BERT [29] for radiology reports. The vision and text encoders are initialized from publicly available pretrained weights, while the EHR encoder is trained from scratch. All components are trained end-to-end after a shared CXR encoder warm start (Appendix F). The representation predictor g_ϕ is implemented as a multi-layer Transformer with hidden dimension 256, and the mask predictor h_ω is a scoring head on the outputs of its EMA copy $g_{\bar{\phi}}$. The fusion module f_ψ is a single-layer Transformer that aggregates multimodal representations for final prediction.

Training Configuration. We train all models end-to-end with Adam and early stopping. Loss weights, learning rates, EMA decay, and other training details are reported in Appendix F, and the default mask ratios in Appendix D.10.

Missing Modality Protocol. On the fully paired subset, we simulate missing scenarios during training. Each training batch is composed of equal proportions (1/3 each) of three conditions: all modalities present, exactly one modality missing, and exactly two modalities missing. For evaluation at missing ratio r, a fraction r of test samples has missing modalities, split equally between partial missing (one modality absent) and severe missing (two modalities absent) scenarios, and the rest are

Table 1: Performance comparison across three clinical tasks at 50% missing modality ratio on Multimodal MIMIC Benchmark. We report mean±std across five runs. †Excluded from LoS due to discrete label requirements. Best in **bold**, second best underlined.

Method	In-hospital Mortality		Phenotyping		Length-of-Stay	
	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	MAE (↓)	κ (↑)
Missing modality methods						
SMIL	78.4±1.7	43.4±4.9	70.5±0.4	41.6±0.3	76.4±0.8	23.7±2.3
MUSE†	77.5±1.4	44.8±2.1	68.7±0.5	39.8±0.6	–	–
ShaSpec	77.6±4.1	42.5±7.2	68.0±1.0	38.0±1.3	77.8±0.3	18.6±6.9
M3Care	78.1±1.1	46.8±1.4	69.0±0.4	39.3±1.1	74.8±1.3	26.1±1.9
DRIM	78.7±1.6	47.7±4.0	69.3±0.6	41.1±1.4	76.2±1.6	25.7±1.8
DiCMoR†	77.3±1.0	47.1±2.5	70.0±0.0	42.2±0.2	–	–
FuseMoE	78.6±0.6	47.6±2.1	70.5±0.4	42.5±0.3	75.4±2.1	27.7±0.5
Flex-MoE	79.6±1.2	48.5±2.8	69.5±0.7	40.2±0.7	75.7±1.8	26.8±3.7
SimMLM	78.6±1.2	46.9±1.5	64.8±0.5	34.7±1.2	88.8±5.9	16.0±6.0
Robult	78.5±2.3	44.5±2.9	67.6±0.5	38.2±0.8	76.7±0.8	20.7±4.7
MMVM-VAE	79.8±0.3	47.8±0.5	68.9±0.7	38.0±1.0	78.1±0.2	23.5±0.8
Cross-modal representation learning						
M3-JEPA	76.4±2.2	43.7±2.4	64.2±0.4	34.2±0.8	83.5±1.1	17.2±3.3
CEPA (Ours)	**81.1**±0.6	**50.6**±2.0	**70.8**±0.3	**42.8**±0.4	**72.9**±0.7	**30.2**±3.6

Table 2: Ablation on loss components (mortality, 50% missing).

$\mathcal{L}_{\text{obs}}$	$\mathcal{L}_{\text{cross}}$	$\mathcal{L}_{\text{CF}}$	AUROC (↑)	AUPRC (↑)
✔	✗	✗	78.8±0.8	43.7±1.3
✔	✔	✗	79.5±0.7	46.2±1.5
✔	✗	✔	80.3±0.9	48.1±1.8
✗	✔	✔	80.5±0.7	48.8±1.6
✔	✔	✔	**81.1**±0.6	**50.6**±2.0

Table 3: Ablation on masking strategies (mortality, 50% missing).

Masking Strategy	AUROC (↑)	AUPRC (↑)
No Masking	79.2±0.9	45.3±1.6
Random Masking	79.6±0.8	46.5±1.5
Block Masking	79.9±0.7	47.2±1.4
Learned (no EMA)	79.5±1.0	46.0±1.5
Learned (no cross-modal)	80.6±0.8	48.8±1.4
Cross-modal Masking (Ours)	**81.1**±0.6	**50.6**±2.0

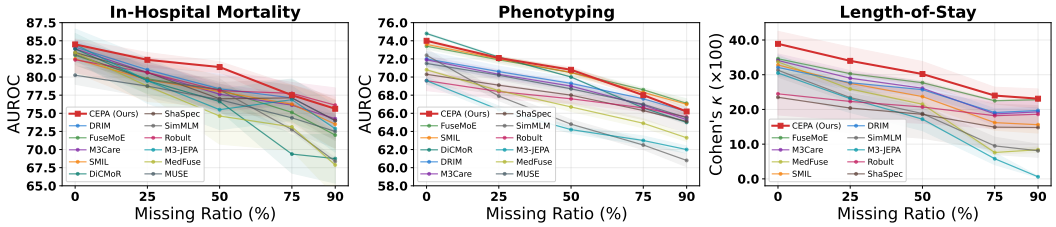


Figure 4: Performance degradation across missing ratios (0%–90%) on the multimodal MIMIC clinical tasks. CEPA retains the most consistent advantage in the practical 25–50% regime, while at 90% missing all methods converge toward a single-modality information bottleneck.

complete. Main results use $r = 50\%$. On the partially paired cohort, models receive the modalities that are naturally available for each patient without applying simulated modality drops. All results are averaged over five runs with different seeds. Full protocol details are in Appendix E.

3.5 Clinical Prediction Results on Multimodal MIMIC Data

Main results. Table 1 shows that CEPA achieves the best mean performance on all three tasks, and the improvement over the strongest baseline for each metric is statistically significant in a seed-paired test for every metric except length-of-stay κ (paired p-values and 95% confidence intervals in Appendix D.4). Existing methods that model cross-modal relationships, such as FuseMoE and Flex-MoE, are among the stronger baselines, but they do not enforce distributional consistency for predicted missing representations. By using context-conditional CF matching, CEPA yields consistent gains across classification and regression tasks. Additional comparisons with Complete-Data, Zero-Fill, Available-Only, and HAIM-MIMIC-MM baselines are in Appendix D.3 and Appendix D.6.

Robustness under severe missing ratios. Figure 4 shows that CEPA maintains strong performance up to 50% missing, where context-conditional CF alignment yields the most consistent advantage. At 90% missing, all methods converge toward a single-modality information bottleneck and the gap narrows.

Table 4: Evaluation under natural MIMIC missingness on in-hospital mortality task.

Method	AUROC (↑)
DRIM	83.3±0.2
FuseMoE	83.6±0.6
Flex-MoE	84.4±0.4
CEPA (Ours)	**85.1**±0.5

Natural missingness. Unlike the controlled protocol, where modalities are dropped at random, this cohort preserves the modality availability observed in the data. Because TXT is the radiology report of the paired CXR study, the two are jointly present or absent, giving two observation patterns: EHR-only (roughly 74% of ICU stays) and all three modalities (EHR, CXR, and TXT). This availability pattern may reflect clinical ordering decisions and institutional workflow and is not generated by our random-removal mechanism. Table 4 reports in-hospital mortality in this setting, where CEPA outperforms DRIM, FuseMoE, and Flex-MoE.

Table 5: Non-clinical generalization on CMU-MOSEI, averaged over the seven observed-modality subsets (Avg7) and five seeds. Best in **bold**, second best underlined.

Method	Avg7 Acc-2 (↑)	Avg7 F1 (↑)
COMPLETER	74.3±0.9	73.7±0.8
DiCMoR	77.0±0.4	76.1±1.4
IMDer	75.8±0.9	75.2±0.7
MoMKE	78.2±0.3	77.6±0.3
MMVM-VAE	78.1±0.2	76.9±0.3
SimMLM	78.0±0.3	77.6±0.8
CMAD	78.2±0.4	77.5±0.6
CEPA (Ours)	**78.7**±0.3	**78.1**±0.5

Masking strategy ablation. Random and block masking improve over no masking (Table 3), but our cross-modal masking yields the largest gain (+1.9 AUROC, +5.3 AUPRC), suggesting that data-driven masks help capture meaningful inter-modal dependencies. Removing EMA guidance or cross-modal conditioning from the learned selection reduces this gain.

Loss ablation. Table 2 shows that adding $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{CF}}$ progressively improves performance, with the full objective performing best. The larger gain from $\mathcal{L}_{\text{CF}}$ indicates that context-conditional distribution alignment provides stronger supervision than pointwise cross-modal reconstruction. Additional alignment and representation diagnostics are in Appendix D.8 and Appendix D.1.

Additional analyses. Qualitative results for learned masks and predicted representations, along with computational cost measurements, are provided in Appendix D.

3.6 Multimodal Sentiment Prediction Results on CMU-MOSEI

We also evaluate CEPA on CMU-MOSEI, a multimodal sentiment benchmark with audio, language, and vision modalities, under the common incomplete-modality protocol over every subset of observed modalities, across five seeds. All methods use the same feature cache, encoder initialization, data split, and missing-pattern assignments. Table 5 reports the average over the seven observed-modality subsets (Avg7). CEPA attains the highest Avg7 on both non-zero binary accuracy (Acc-2) and weighted F1, ahead of strong recent baselines such as MoMKE, CMAD, and MMVM-VAE. Per-subset results and the full experimental setup are in Appendix D.5.

4 Conclusion

We presented CEPA, a missing-modality learning framework that combines adaptive cross-modal masking, embedding-level prediction, and context-conditional distribution alignment to mitigate representation collapse without raw-input reconstruction. CEPA achieves comparable or better performance than strong missing-modality baselines on three MIMIC clinical tasks and on the CMU-MOSEI sentiment benchmark. At the same time, CEPA optimizes task-relevant latent prediction rather than faithful recovery of missing inputs, and is therefore unsuitable when human-interpretable reconstruction is required. Broader external cohorts and additional domains remain promising directions for future work.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST), No. RS-2022-II220713, Meta-learning Applicable to Real-world Problems, and No. RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

References

[1] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[2] Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 5299–5318, 2024.

[3] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.

[4] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2418–2428, 2022.

[5] 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, pages 16416–16424, 2024.

[6] 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, pages 2302–2310, 2021.

[7] Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24068–24077, 2025.

[8] Maria Boyko, Aleksandra Beliaeva, Dmitriy Kornilov, Alexander Bernstein, and Maxim Sharaev. impuTMAE: Multi-modal transformer with masked pre-training for missing modalities imputation in cancer survival prediction. *arXiv preprint arXiv:2508.09195*, 2025.

[9] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887, 2023.

[10] Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. DRIM: Learning disentangled representations from incomplete multimodal healthcare data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 163–173. Springer, 2024.

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

[12] 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*, pages 16000–16009, 2022.

[13] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[14] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2025.

[15] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[16] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *International Conference on Machine Learning*, pages 2566–2585, 2024.

[17] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[18] R. Balestriero and Y. LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

[19] 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, 2023.

[20] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR database. *PhysioNet10*, 13026:C2JT1Q, 2024.

[21] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[22] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034, 2023.

[23] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. FuseMoE: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37: 67850–67900, 2024.

[24] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37: 98782–98805, 2024.

[25] Duy A Nguyen, Abhi Kamboj, and Minh N Do. Robult: Leveraging redundancy and modality-specific features for robust multimodal learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5985–5993, 2025.

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

[27] Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture. In *International Conference on Machine Learning*. PMLR, 2025. URL https://proceedings.mlr.press/v267/lei25b.html.

[28] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[29] 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*, pages 1–21. Springer, 2022.

[30] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[31] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446, 2024.

[32] 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 (CVPR)*, pages 18177–18186, 2022.

[33] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 74–82. Springer, 2019.

[34] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39:1170–1183, 2019.

[35] Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4626–4636, 2025.

[36] He Zhao, Ke Sun, Amir Dezfouli, and Edwin V. Bonilla. Transformed distribution matching for missing value imputation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42159–42186. PMLR, 2023.

[37] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11174–11183, June 2021.

[38] Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

[39] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *Computer Vision – ECCV 2022*, volume 13697 of *Lecture Notes in Computer Science*, pages 348–367. Springer, 2022.

[40] 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*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 2022.

[41] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

[42] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022.

[43] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *Computer Vision – ECCV 2022*, volume 13690 of *Lecture Notes in Computer Science*, pages 300–318. Springer, 2022.

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

A Theoretical Analysis and Proofs

Here, we prove **Proposition 2.1**, ensuring that our loss function is bounded and provides stable gradients, preventing the explosion issues common in distribution matching.

Proposition A.1 (Gradient Boundedness)**.** *Let $\|\mathbf{a}_\ell\|_2 = 1$ be unit projection directions. The gradient of $\mathcal{L}_{CF}^{(m)}$ with respect to the prediction $\widehat{\mathbf{h}}$ is bounded in terms of the sampled frequencies. Specifically, $\|\nabla_{\widehat{\mathbf{h}}}\mathcal{L}_{CF}^{(m)}\|_2 \leq 4 \max_k |\tau_k|$.*

Proof. The proof consists of two parts: the boundedness of the loss value and of its gradient.

1. Boundedness of Loss. Both the predicted and target characteristic functions are kernel-weighted averages of complex exponentials on the unit circle. For the target CF over observed samples B_m:

$$|\varphi_{\mathbf{c}_i}^{(m)}| = \left| \sum_{j \in B_m} w_{ij}^{(m)} e^{\mathrm{i}\tau \mathbf{a}^\top \mathbf{h}_j^{(m)}} \right| \leq \sum_{j \in B_m} w_{ij}^{(m)} |e^{\mathrm{i}\theta}| = 1. \tag{15}$$

Similarly, for the predicted CF over missing samples $B_m^c := \mathcal{B} \setminus B_m$:

$$|\widehat{\varphi}_{\mathbf{c}_i}^{(m)}| = \left| \sum_{i' \in B_m^c} \widehat{w}_{ii'}^{(m)} e^{\mathrm{i}\tau \mathbf{a}^\top \widehat{\mathbf{h}}_{i'}^{(m)}} \right| \leq \sum_{i' \in B_m^c} \widehat{w}_{ii'}^{(m)} = 1, \tag{16}$$

where $\widehat{w}_{ii'}^{(m)} = \kappa(\mathbf{c}_i, \mathbf{c}_{i'}) / \sum_{i''} \kappa(\mathbf{c}_i, \mathbf{c}_{i''})$ are the self-similarity weights among missing samples. By the triangle inequality, the maximum distance is $|\widehat{\varphi}_{\mathbf{c}_i}^{(m)} - \varphi_{\mathbf{c}_i}^{(m)}| \leq 2$. Therefore, the squared loss is universally bounded: $\mathcal{L}_{\text{CF}}^{(m)} \leq 4$.

2. Gradient Control via Stop-Gradient. Let $\xi_i := \mathbf{a}^\top \widehat{\mathbf{h}}_i^{(m)}$ and $\Delta := \widehat{\varphi}_{\mathbf{c}_i}^{(m)} - \varphi_{\mathbf{c}_i}^{(m)}$. The gradient of the predicted CF has two potential pathways:

$$\frac{\partial \widehat{\varphi}_{\mathbf{c}_i}^{(m)}}{\partial \widehat{\mathbf{h}}_i^{(m)}} = \sum_{i'} \underbrace{\frac{\partial \widehat{w}_{ii'}^{(m)}}{\partial \widehat{\mathbf{h}}_i^{(m)}}}_{\text{kernel path}} e^{\mathrm{i}\tau \xi_{i'}} + \sum_{i'} \widehat{w}_{ii'}^{(m)} \cdot \mathrm{i}\tau \mathbf{a} \cdot \mathbf{1}[i' = i] \cdot e^{\mathrm{i}\tau \xi_{i'}}. \tag{17}$$

The first term vanishes because the weights depend on $\widehat{\mathbf{h}}_i^{(m)}$ only through the contexts $\mathbf{c}$, which are computed from observed embeddings. We additionally apply stop-gradient to the kernel weights, i.e., $w_{ij}^{(m)} \leftarrow \mathsf{stopgrad}(w_{ij}^{(m)})$ and $\widehat{w}_{ii'}^{(m)} \leftarrow \mathsf{stopgrad}(\widehat{w}_{ii'}^{(m)})$, so that gradients do not flow into the contexts, which would otherwise let the model manipulate context representations to trivially minimize the loss. With stop-gradient:

$$\begin{aligned} \frac{\partial \widehat{\varphi}_{\mathbf{c}_i}^{(m)}}{\partial \widehat{\mathbf{h}}_i^{(m)}} &= \sum_{i'} \underbrace{\frac{\partial \widehat{w}_{ii'}^{(m)}}{\partial \widehat{\mathbf{h}}_i^{(m)}}}_{=0 \text{ (stop-grad)}} e^{\mathrm{i}\tau \xi_{i'}} + \sum_{i'} \widehat{w}_{ii'}^{(m)} \cdot \mathrm{i}\tau \mathbf{a} \cdot \underbrace{\mathbf{1}[i' = i]}_{=1 \text{ only if } i'=i} \cdot e^{\mathrm{i}\tau \xi_{i'}} \\ &= \widehat{w}_{ii}^{(m)} \cdot \mathrm{i}\tau \mathbf{a} \cdot \exp(\mathrm{i}\tau \xi_i), \end{aligned}$$

where $\widehat{w}_{ii}^{(m)} \leq 1$ is the self-weight of sample i. Taking the norm of the total gradient:

$$\begin{aligned} \|\nabla_{\widehat{\mathbf{h}}_i^{(m)}} |\Delta|^2\|_2 &= \|2\Re(\overline{\Delta} \cdot \widehat{w}_{ii}^{(m)} \cdot \mathrm{i}\tau \mathbf{a} \cdot e^{\mathrm{i}\tau \xi_i})\|_2 \\ &\leq 2 \cdot \underbrace{|\Delta|}_{\leq 2} \cdot \underbrace{\widehat{w}_{ii}^{(m)}}_{\leq 1} \cdot |\tau| \cdot \underbrace{\|\mathbf{a}\|_2}_{1} \cdot \underbrace{|e^{\mathrm{i}\tau \xi_i}|}_{1} \\ &\leq 4|\tau| \leq 4 \max_k |\tau_k|. \end{aligned}$$

The kernel weighting introduces an additional factor $\widehat{w}_{ii}^{(m)} \leq 1$, which can only *reduce* the gradient magnitude compared to the point-mass case. The gradient norm is linear in the largest sampled frequency $\max_k |\tau_k|$ and independent of the context distribution or kernel bandwidth, so the gradient cannot explode during optimization.

Algorithm 1 CEPA Training

Require: Dataset $\mathcal{D}$, encoders $\{\mathcal{E}^{(m)}\}$, predictor g_ϕ, mask head h_ω, fusion f_ψ
Require: Loss weights α, β, γ; EMA decay λ
Initialize $\bar{\phi} \leftarrow \phi$ EMA predictor
for each batch $(\mathbf{X}, y)$ **do**
  Sample missing set $\mathcal{M}$; observed set $\mathcal{O} \leftarrow [M] \setminus \mathcal{M}$
  $\mathbf{Z}^{(m)} \leftarrow \mathcal{E}^{(m)}(\mathbf{X}^{(m)})$ for $m \in \mathcal{O}$
  // Cross-modal Masking (Section 2.3)
  $\mathbf{M}^{(m)} \leftarrow \texttt{TopK}\big(h_\omega([g_{\bar{\phi}}(\mathbf{Z}^{\mathcal{O}}, \mathbf{U}^{\mathcal{M}})]^{(m)}), K^{(m)}\big)$ for $m \in \mathcal{O}$ Eq. (3)
  $\tilde{\mathbf{Z}}^{(m)} \leftarrow \mathcal{E}^{(m)}(\mathbf{X}^{(m)} \odot (1 - \mathbf{M}^{(m)}) + \mathbf{V}^{(m)} \odot \mathbf{M}^{(m)})$ Eq. (1)
  $\mathbf{M}'^{(m)}, \tilde{\mathbf{Z}}'^{(m)}$: same as above with $K'^{(m)}$
  // Cross-modal Prediction (Section 2.2)
  $\widehat{\mathbf{Z}} \leftarrow g_\phi(\tilde{\mathbf{Z}}^{\mathcal{O}}, \mathbf{U}^{\mathcal{M}})$ Eq. (2)
  $\mathcal{L}_{\text{obs}} \leftarrow \|(\widehat{\mathbf{Z}}^{(m)} - \mathbf{Z}^{(m)}) \odot \mathbf{M}^{(m)}\|_F^2$ for $m \in \mathcal{O}$ Eq. (4)
  // Cross-modal Consistency (Section 2.3)
  $\widehat{\widehat{\mathbf{Z}}} \leftarrow g_\phi(\{\tilde{\mathbf{Z}}'^{(m)}\}_{m \in \mathcal{O}}, \widehat{\mathbf{Z}}^{\mathcal{M}})$ Eq. (5)
  $\mathcal{L}_{\text{cross}} \leftarrow \|(\widehat{\widehat{\mathbf{Z}}}^{(m)} - \mathbf{Z}^{(m)}) \odot \mathbf{M}'^{(m)}\|_F^2$ for $m \in \mathcal{O}$ Eq. (6)
  // CF Alignment & Task (Section 2.4)
  $\mathcal{L}_{\text{CF}} \leftarrow$ CF matching loss Eq. (12)
  $\hat{y} \leftarrow f_\psi(\mathbf{h}^{\mathcal{O}}, \widehat{\mathbf{h}}^{\mathcal{M}})$; $\mathcal{L}_{\text{task}} \leftarrow \text{Loss}(\hat{y}, y)$
  // Update
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{obs}} + \beta\mathcal{L}_{\text{cross}} + \gamma\mathcal{L}_{\text{CF}}$ Eq. (14)
  Update ϕ, ψ, ω; $\bar{\phi} \leftarrow \lambda\bar{\phi} + (1 - \lambda)\phi$
end for

The argument above bounds the gradient of a single query's residual $|\Delta_i|^2$ with respect to its *own* prediction $\widehat{\mathbf{h}}_i^{(m)}$, through the self-weight $\widehat{w}_{ii}^{(m)}$. For the batch-averaged loss

$$\mathcal{L}_{\text{CF}}^{(m)} = \frac{1}{|B_m^c|} \sum_{q \in B_m^c} \frac{1}{N_a N_\tau} \sum_{\ell,k} \big|\Delta_q(\tau_k, \mathbf{a}_\ell)\big|^2, \qquad \Delta_q := \widehat{\varphi}_{\mathbf{c}_q}^{(m)} - \varphi_{\mathbf{c}_q}^{(m)}, \tag{18}$$

however, each prediction $\widehat{\mathbf{h}}_i^{(m)}$ also enters the predicted-side estimator of every other query q, since $\widehat{\varphi}_{\mathbf{c}_q}^{(m)} = \sum_{i' \in B_m^c} \widehat{w}_{qi'}^{(m)} e^{\mathrm{i}\tau\mathbf{a}^\top \widehat{\mathbf{h}}_{i'}^{(m)}}$ contains the term $i' = i$ with weight $\widehat{w}_{qi}^{(m)}$. By the same unit-circle argument as in Part 2, the gradient of query q's residual with respect to $\widehat{\mathbf{h}}_i^{(m)}$ is bounded by $4 \max_k |\tau_k| \, \widehat{w}_{qi}^{(m)}$. Summing these cross-query contributions and averaging over the $|B_m^c|$ queries gives the per-prediction bound

$$\big\|\nabla_{\widehat{\mathbf{h}}_i^{(m)}} \mathcal{L}_{\text{CF}}^{(m)}\big\|_2 \;\le\; \frac{4 \max_k |\tau_k|}{|B_m^c|} \, C_i, \qquad C_i := \sum_{q \in B_m^c} \widehat{w}_{qi}^{(m)}. \tag{19}$$

Because the weights are normalized over the predicted side, $\sum_{i \in B_m^c} \widehat{w}_{qi}^{(m)} = 1$ for each q, and hence $\sum_{i \in B_m^c} C_i = \sum_{q \in B_m^c} \sum_{i \in B_m^c} \widehat{w}_{qi}^{(m)} = |B_m^c|$. Summing the per-prediction bound over all predictions therefore yields

$$\sum_{i \in B_m^c} \big\|\nabla_{\widehat{\mathbf{h}}_i^{(m)}} \mathcal{L}_{\text{CF}}^{(m)}\big\|_2 \;\le\; \frac{4 \max_k |\tau_k|}{|B_m^c|} \sum_{i \in B_m^c} C_i \;=\; 4 \max_k |\tau_k|. \tag{20}$$

Accounting for the cross-query contributions thus does not enlarge the bound: the total gradient over all predictions in $\mathcal{L}_{\text{CF}}^{(m)}$ remains controlled by the same constant $4 \max_k |\tau_k|$. □

B Training Pipeline

Algorithm 1 summarizes the training procedure. Each phase corresponds to the method sections: adaptive masking generates context-aware masks via EMA predictor, cross-modal prediction reconstructs both masked observed and missing modalities, and CF alignment regularizes the predicted representations to match the conditional distribution of observed ones.

C Related Work

C.1 Modeling under Missing Modalities

Missing modalities remain a key challenge in deploying multimodal systems across clinical, affective, and perceptual tasks [5, 4, 30, 31, 6, 9]. Even multimodal Transformers are sensitive to missing modalities, and the best fusion strategy depends on the dataset [32]. Early work used variational or adversarial generators to recover missing views [33, 34], but accurately synthesizing one modality from another, such as medical images from EHR data, can be infeasible [5]. Recent methods instead reason in latent space through diverse strategies. Disentanglement-based approaches decompose representations into shared and modality-specific factors [9, 5, 10]. Meta-learning and distillation methods transfer knowledge from complete-modality data, via Bayesian meta-learned feature reconstruction [6] or correlation-aware teacher–student distillation [35]. Graph-based methods propagate information over patient-similarity graphs [4] or patient–modality bipartite graphs with contrastive objectives [11]. Generative approaches recover missing modalities with class-conditional normalizing flows [22] or score-based diffusion [30] so that recovered data follow the original distribution. Mixture-of-experts approaches dynamically combine modality-specific experts or representations through learned routing and gating [31, 23, 24, 7]. Beyond multimodal learning, missing-value imputation has also been cast as distribution matching between batches mapped to a latent space with invertible networks [36]. Predicting missing representations from observed ones is itself common, through dual prediction [37], normalizing flows [22], or diffusion [30], and these methods derive their recovery or generative supervision from complete samples in which the missing modality is observed. M3Care instead imputes a point estimate aggregated from similar patients [4]. CEPA differs in its training target, which is the distribution of representations observed in similar contexts.

C.2 Representation Learning for Multimodal Robustness

Self-supervised learning (SSL) leverages unlabeled data to capture modality-agnostic structure, complementing supervised adaptation. Contrastive and distribution-distance objectives align the representations of modality-missing and modality-complete data [38] or enforce mutual consistency to prevent modality collapse [11], while Robult combines a soft positive-unlabeled contrastive loss with latent reconstruction to retain redundant and modality-specific information [25]. Masked autoencoding treats missing modalities as extreme masks: MultiMAE masks tokens across input modalities for cross-modal predictive coding [39], and impuTMAE extends ViT-MAE to heterogeneous, incomplete medical inputs [8]. Balestriero and LeCun show, however, that reconstruction biases features toward high-variance but perceptually uninformative subspaces, which masking only partly mitigates [16]. Latent-prediction methods instead predict representations rather than inputs: data2vec predicts latent targets of the full input from a masked view using an EMA teacher [40], I-JEPA predicts target-block embeddings from a context block [13], and M3-JEPA predicts the embedding of one modality from another to align them [27]. Such joint-embedding objectives risk complete or dimensional collapse, the latter occurring even in contrastive learning [41], and SIGReg counters this by constraining embeddings toward an isotropic Gaussian [18]. Informed (non-random) masking has also been studied in SSL, with a masking function learned adversarially [42] or masks guided by the attention of an EMA teacher [43].

D Additional Experimental Results

D.1 Representation Diagnostics and Qualitative Analysis

Downstream performance alone does not prove that predicted missing representations preserve meaningful structure. We therefore measure collapse-related diagnostics on MIMIC complete-case test samples by masking one modality and comparing predicted representations with the corresponding real representations.

Qualitative diagnostics. We also visualize the learned masks and the imputed representations. Predicted CXR masks change when TXT context is available (Figure 5), and imputed fused representations align with their real counterparts rather than collapsing to a single fallback region (Figure 6). Together with Table 6, this indicates context-dependent latent completion rather than a static token.

Table 6: Representation diagnostics on MIMIC mortality under one-modality-missing evaluation. Higher variance and diversity indicate less collapse.

Configuration	Repr. Variance (↑)	Diversity (↑)
No $\mathcal{L}_{CF}$	0.466	0.488
With $\mathcal{L}_{CF}$	**0.540**	**0.568**

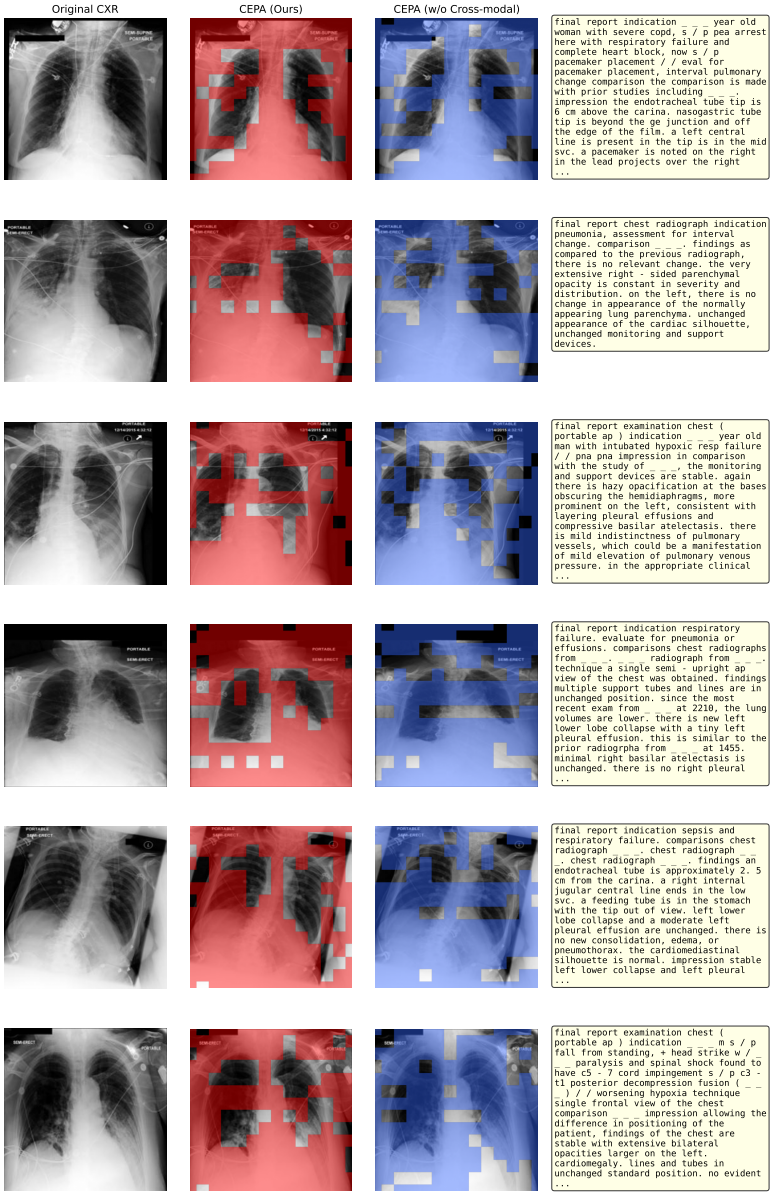


Figure 5: Qualitative mask predictions on MIMIC mortality test cases. Each row shows the original CXR (left), CEPA's predicted mask (center, red overlay), the predicted mask without cross-modal conditioning on TXT (right, blue overlay), and the paired radiology report (far right). Cross-modal conditioning preserves clinically salient thoracic regions while masking peripheral areas, whereas removing cross-modal context shifts the mask pattern to less informative regions.

D.2 Computational Cost

We compare inference-time cost under the same encoder backbone. Masking, reconstruction, and CF alignment are training-time objectives and are inactive at inference.

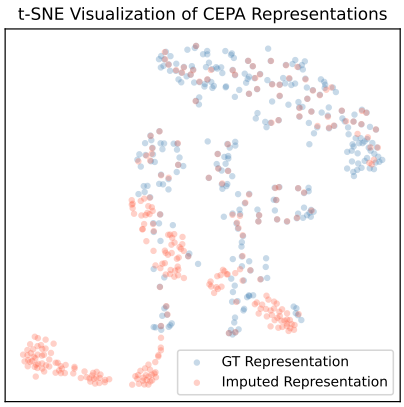


Figure 6: t-SNE visualization of pooled fused representations on the MIMIC mortality test set. Ground-truth representations (blue) and CEPA-imputed representations (red).

Table 7: Computational cost on MIMIC mortality using NVIDIA RTX PRO 6000, batch size 16, and three measured inference passes after warmup.

Method	Params (M)	Infer (samples/s) (↑)
SMIL	205.8	173.2
ShaSpec	206.4	172.0
CEPA	213.8	171.9
FuseMoE	210.2	169.1
Flex-MoE	212.5	144.9
DRIM	417.3	78.7

CEPA adds about 4% parameters over SMIL and ShaSpec, the lightest baselines, and has comparable inference throughput. The main overhead is training-time only, where the additional prediction and alignment objectives increase wall-clock time by approximately 1.8×.

D.3 Performance Bounds and Available-Only Baseline

We add explicit empirical anchors for the missing-modality setting. The complete-data model provides an approximate upper bound, while static replacement and available-only prediction provide simple lower-cost alternatives. This comparison tests whether latent completion contributes predictive value beyond simply aggregating the observed modalities.

Table 8: Performance bounds on MIMIC mortality at 50% missing modality ratio. AUROC is reported as mean±std. Recovery rate is computed relative to Zero Fill and Complete Data.

Method	AUROC (↑)	Recovery Rate (↑)
Complete Data (0% missing ceiling)	86.7	100.0%
Zero Fill	77.5 ±0.8	0.0%
Available-Only	78.2 ±0.7	7.6%
DRIM	78.7 ±1.6	13.0%
FuseMoE	78.6 ±0.6	12.0%
Flex-MoE	79.6 ±1.2	22.8%
MMVM-VAE	79.8 ±0.3	25.0%
CEPA	**81.1 ±0.6**	**39.1%**

CEPA outperforms Available-Only by 2.9 AUROC, indicating that the predicted latent representations provide task-useful information beyond directly using the observed modalities alone. The result also separates CEPA from zero-cost static replacement strategies.

D.4 Paired Significance Tests and Confidence Intervals

To assess whether the mean improvements in Table 1 are statistically reliable, we run seed-paired tests at 50% missingness. For each metric we compare CEPA against the baseline with the strongest mean performance on that metric, using matched seeds, and report the paired p-value together with the 95% confidence interval of the per-seed difference (Table 9).

Table 9: CEPA versus the strongest baseline per metric at 50% missingness. Phenotyping AUROC lists both baselines that tie for strongest. Δ is the per-seed difference (CEPA minus baseline) and the interval is its 95% confidence interval.

Task	**Metric**	**CEPA**	**Strongest baseline**	**Paired p**	**Δ 95% CI**
In-hospital mortality	AUROC	81.1 ±0.6	MMVM-VAE 79.8 ±0.3	0.018	$[+0.37, +2.23]$
	AUPRC	50.6 ±2.0	Flex-MoE 48.5 ±2.8	0.003	$[+1.17, +3.03]$
Phenotyping	AUROC	70.8 ±0.3	SMIL 70.5 ±0.4	0.009	$[+0.12, +0.48]$
	AUROC	70.8 ±0.3	FuseMoE 70.5 ±0.4	0.023	$[+0.07, +0.53]$
	AUPRC	42.8 ±0.4	FuseMoE 42.5 ±0.3	0.013	$[+0.10, +0.50]$
Length-of-stay	MAE	72.9 ±0.7	M3Care 74.8 ±1.3	0.002	$[-2.65, -1.15]$
	$\kappa \times 100$	30.2 ±3.6	FuseMoE 27.7 ±0.5	0.150	$[-1.37, +6.37]$

For every metric except length-of-stay κ, the paired difference favors CEPA and the 95% confidence interval lies entirely on the favorable side of zero. On length-of-stay κ the interval includes zero, so we do not claim a significant improvement on that metric.

D.5 Non-clinical Evaluation on CMU-MOSEI

We evaluate on CMU-MOSEI, an audio–language–vision sentiment benchmark, as a non-clinical test of generalization. The processed set contains $3{,}225$ videos and $22{,}856$ utterances, split into $2{,}249/300/676$ train/validation/test videos. We follow the standard pre-extracted-feature setting: 512-dimensional Wav2Vec acoustic (A), $1{,}024$-dimensional DeBERTa language (L), and $1{,}024$-dimensional MANet visual (V) features, with upstream encoders not fine-tuned. Each utterance feature is projected to 256 dimensions, and each video is a variable-length sequence of utterance tokens encoded by a lightweight modality-specific temporal Transformer. All compared methods share the same feature cache, projection and temporal-encoding interface, encoder initialization, data split, and missing-pattern assignments; their method-specific recovery and fusion modules operate on the resulting representations.

During training we remove one modality with probability $2/3$ and two modalities with probability $1/3$ at the video level. We train each method five times with seeds 0–4, select the checkpoint with the lowest validation MAE, and test on all seven observed-modality configurations (ALV, AL, AV, LV, A, L, V). Avg7 denotes the average over these seven configurations. Following the standard protocol we report non-zero-label binary accuracy (Acc-2) and weighted F1. Baselines span several paradigms: expert-based learning (MoMKE [31], SimMLM [7]), distribution-consistent flow recovery (DiCMoR [22]), diffusion-based recovery (IMDer [30]), teacher–student distillation (CMAD [35]), variational generative recovery (MMVM-VAE [26]), and contrastive multi-view completion (COMPLETER [37]).

CEPA obtains the highest Avg7 for both Acc-2 and weighted F1. It also attains the highest Acc-2 under AV and A, where the language modality is unavailable, so its advantage is not confined to configurations that retain the dominant language modality.

D.6 Evaluation on HAIM-MIMIC-MM

To complement our MedFuse-based benchmark, we additionally evaluate on HAIM-MIMIC-MM, a standardized multimodal MIMIC benchmark. HAIM and our main setup serve different purposes: HAIM provides a community-recognized construction for reproducibility, while our MedFuse-based cohort enables evaluation under naturally observed modality absence. We therefore use HAIM as an additional benchmark check rather than as external-domain evidence.

Table 10: Per-subset **non-zero binary accuracy (Acc-2)** on CMU-MOSEI over five seeds (mean±std). A: audio, L: language, V: vision. Avg7 averages the seven subsets. Per-column best in **bold**.

Method	ALV	AL	AV	LV	A	L	V	Avg7
COMPLETER (CVPR'21)	80.1±0.9	80.8±0.2	66.2±2.3	80.4±0.4	66.4±1.3	81.1±0.2	64.7±1.7	74.3±0.9
DiCMoR (ICCV'23)	84.9±0.5	85.0±0.7	66.3±1.1	84.8±1.0	66.1±2.0	84.7±1.1	66.9±1.2	77.0±0.4
IMDer (NeurIPS'23)	84.4±0.4	84.1±0.6	65.9±2.0	83.9±1.0	64.3±1.7	83.9±1.3	64.0±1.8	75.8±0.9
MoMKE (ACM MM'24)	86.3±0.2	**86.3**±0.1	68.7±0.7	85.7±0.2	67.7±1.3	85.4±0.2	66.9±1.0	78.2±0.3
MMVM-VAE (NeurIPS'24)	85.5±0.3	85.7±0.4	69.2±0.4	85.5±0.3	68.0±0.1	85.4±0.2	**67.3**±0.2	78.1±0.2
SimMLM (ICCV'25)	85.7±0.2	85.7±0.3	69.4±1.1	85.9±0.2	68.5±1.0	85.5±0.4	65.6±1.3	78.0±0.3
CMAD (ICCV'25)	86.2±0.2	86.2±0.3	68.7±1.1	86.1±0.4	68.4±0.8	85.9±0.4	65.9±1.4	78.2±0.4
CEPA (Ours)	**86.4**±0.2	**86.3**±0.3	**70.0**±0.9	**86.5**±0.6	**69.4**±0.5	**86.3**±0.2	66.0±0.7	**78.7**±0.3

Table 11: Per-subset **weighted F1** on CMU-MOSEI over five seeds (mean±std). Columns and conventions as in Table 10.

Method	ALV	AL	AV	LV	A	L	V	Avg7
COMPLETER (CVPR'21)	80.1±0.8	80.9±0.2	65.8±2.0	80.4±0.4	64.1±1.5	81.2±0.2	63.7±2.6	73.7±0.8
DiCMoR (ICCV'23)	84.8±0.3	84.9±0.5	65.2±1.7	84.8±0.8	64.0±2.2	84.6±0.9	64.1±2.6	76.1±1.4
IMDer (NeurIPS'23)	84.2±0.7	84.1±0.5	64.7±2.1	83.7±0.6	63.2±1.3	83.0±1.2	63.2±1.3	75.2±0.7
MoMKE (ACM MM'24)	86.2±0.2	**86.2**±0.1	67.4±0.4	85.6±0.3	67.4±0.5	85.3±0.3	65.0±0.4	77.6±0.3
MMVM-VAE (NeurIPS'24)	85.4±0.3	85.6±0.4	67.1±0.7	85.4±0.3	63.8±0.2	85.4±0.1	**65.3**±0.3	76.9±0.3
SimMLM (ICCV'25)	85.5±0.3	85.6±0.3	**68.3**±2.3	85.8±0.2	**68.3**±0.8	85.5±0.3	64.6±1.7	77.6±0.8
CMAD (ICCV'25)	86.2±0.2	86.1±0.3	67.3±1.3	86.1±0.3	65.8±1.3	85.9±0.3	65.0±2.6	77.5±0.6
CEPA (Ours)	**86.3**±0.2	**86.2**±0.3	68.1±1.4	**86.2**±0.6	**68.3**±0.9	**86.2**±0.2	65.1±1.4	**78.1**±0.5

D.7 Training Missingness Distribution Sensitivity

The main controlled protocol uses a balanced missingness curriculum with equal proportions of complete, one-missing, and two-missing samples. To test whether the method is overly tuned to this curriculum, we compare against a skewed training distribution in which EHR is removed more often than under the balanced curriculum. Both models use the same controlled missing-ratio test protocol.

The performance difference is within 0.3 AUROC across the evaluated missing ratios, suggesting that the learned latent completion is not brittle to the exact balanced curriculum used in the main controlled experiments.

D.8 Alignment Objective and Loss Ablations

We compare the proposed context-conditional CF alignment against marginal MMD and against removing the distributional regularizer. The comparison isolates the value of context-conditioned distribution matching.

The context-conditional CF term improves over both marginal MMD and no distributional alignment. The loss ablation in Table 2 further shows that $\mathcal{L}_{cross} + \mathcal{L}_{CF}$ remains strong even without $\mathcal{L}_{obs}$, while the full objective performs best. The same comparison in the synthetic setting is shown in Figure 3(c), where CF (Ours) consistently outperforms MMD across $M \in \{1, 2, 4, 8\}$ observed modalities.

D.9 Conditional CF Estimator: Bandwidth and Support Reliability

We examine the reliability of the conditional CF estimator, addressing its sensitivity to the kernel bandwidth, the minibatch support used by the estimator, and the sampling variance of the CF target.

Bandwidth sensitivity. We vary the kernel bandwidth σ and evaluate mortality at 50% missingness over five seeds. In addition to fixed bandwidths, we evaluate the per-minibatch median rule, which pools the ℓ_2-normalized query and reference contexts and sets $\sigma_m = \text{median}\{\|\mathbf{c}_i - \mathbf{c}_j\|\}$. The fixed-bandwidth results vary by at most 0.8 AUROC across the range 0.1 to 2.0, and the value $\sigma = 1.0$ used for the main results is within 0.3 AUROC of the best fixed setting (Table 15). The adaptive rule resolves to a modality-dependent bandwidth around 0.25–0.28 (Table 16).

Table 12: HAIM-MIMIC-MM mortality evaluation under missing modality stress tests. AUROC is reported as mean±std.

Method	50% Missing	75% Missing
DRIM	79.2 ±1.1	75.8 ±1.3
FuseMoE	78.5 ±0.9	74.1 ±1.5
Flex-MoE	79.3 ±1.2	76.5 ±1.4
CEPA	**80.6 ±0.8**	**77.3 ±1.0**

Table 13: Training missingness distribution sensitivity on MIMIC mortality (AUROC, mean±std).

Training Distribution	50%	75%	90%
Balanced (1/3 each)	81.1 ±0.6	78.0 ±1.1	76.5 ±1.3
Skewed-EHR	80.8 ±0.6	78.3 ±0.8	76.8 ±0.9

Support and batch-size stability. Keeping $\sigma = 1.0$, we vary the training batch size over five seeds and report, per minibatch, the observed raw support ($|B_m|$), the observed effective support ($N_{\text{eff}}^{\text{obs}} = 1/\sum_j (w_{ij}^{(m)})^2$), and the insufficient-support rate, the fraction of query-containing modality batches where either side has fewer than two references. At batch size 16, used in the main experiments, the observed raw and effective support are 9.0 and 8.2, and only 0.1% of modality batches fail the minimum-support condition.

Controlled CF-target variance. To isolate reference-set sampling noise from representation learning, we fix the trained model, queries, contexts, projection directions, and frequency grid, and resample an observed reference set of size R 200 times per query, measuring the variance of the resulting CF target averaged over queries, projections, and frequencies. Here R is a controlled reference-set size, not the training batch size. Doubling the reference count approximately halves the measured target variance, and the effective support tracks R, indicating that the weights remain distributed across references rather than concentrating on one.

D.10 Masking Strategy and Mask Ratio Sensitivity

Table 3 decomposes the adaptive masking design into learned token selection, EMA guidance, and cross-modal conditioning. This addresses whether gains arise merely from additional capacity or from the proposed masking mechanism.

Table 19 varies the mask ratio of one modality at a time while keeping the other two at their defaults. This is a post-hoc one-factor sensitivity sweep. All main-paper results use a single default mask-ratio configuration held constant across tables, and performance remains above the strongest baselines across the explored ratios. Figure 7 visualizes the per-modality sensitivity curves.

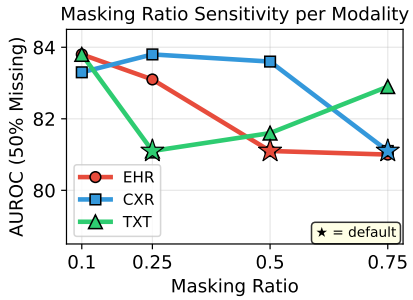


Figure 7: Per-modality masking ratio sensitivity on MIMIC mortality at 50% missing modality ratio. Stars mark the default settings used in the main experiments.

Table 14: Regularization ablation on MIMIC mortality. AUROC is reported as mean±std.

Regularizer	50%	75%	90%
$\mathcal{L}_{obs} + \mathcal{L}_{cross}$ only	79.5 ±0.7	77.1 ±0.9	75.5 ±1.0
MMD	80.4 ±0.8	77.3 ±1.0	75.7 ±1.1
$\mathcal{L}_{CF}$ (Ours)	**81.1 ±0.6**	**78.0 ±1.1**	**76.5 ±1.3**

Table 15: Bandwidth sensitivity on MIMIC mortality at 50% missingness (five seeds).

Bandwidth	M50 AUROC (↑)	M50 AUPRC (↑)
Fixed 0.10	81.4 ±1.4	50.4 ±3.2
Fixed 0.25	81.0 ±1.7	51.4 ±4.0
Fixed 0.50	80.6 ±1.7	50.2 ±3.7
Fixed 0.75	81.4 ±1.2	51.8 ±3.0
Fixed 1.00	81.1 ±0.6	50.6 ±2.0
Fixed 2.00	81.3 ±1.1	51.5 ±2.0
Adaptive (median)	81.7 ±1.5	50.9 ±3.0

Table 16: Adaptive bandwidth at the final epoch (five seeds).

Modality	σ (mean±std)
EHR	0.252 ±0.021
CXR	0.276 ±0.027
TXT	0.278 ±0.038
Overall	0.269 ±0.030

D.11 Per-Missing-Configuration Breakdown

Aggregate missing-ratio evaluation can hide differences across observed modality subsets. We therefore report deterministic missing-configuration results for CEPA on mortality, averaged over five seeds (0–4). Each row lists the modalities observed by the model.

The breakdown shows that the difficulty of latent completion is modality-dependent. Subsets that retain text are strongest, with EHR+TXT and CXR+TXT reaching 84.9 and 84.4 AUROC, while single-modality EHR and CXR are comparable and lower (76.3 and 75.9). This motivates reporting aggregate and per-configuration robustness.

D.12 Limitations and Broader Impacts

CEPA predicts task-relevant latent representations rather than faithful reconstructions of missing inputs, so it is not suited to settings that require human-interpretable recovery. Its context-conditional CF target is formed from a minibatch of observed-side references, so it requires sufficient observed-side support in each minibatch, which a batch size of 16 provides in our setting (Appendix D.9). For the same reason, patients in sparse regions of the context space receive conditional targets built from fewer similar references. Gains are clearest on in-hospital mortality and smaller on phenotyping and length-of-stay. Our evaluation covers the MIMIC data family and one non-clinical benchmark (CMU-MOSEI), and broader external cohorts, together with calibration and subgroup analysis, remain important future work before clinical deployment.

More robust prediction under missing modalities may support clinical decision-making when some tests are unavailable. However, imputed representations are not observations, and a model trained on MIMIC data can inherit its cohort biases and degrade under distribution shift. CEPA should therefore not be used as a standalone clinical tool without prospective validation, calibration, and governance.

E Detailed Experimental Setup

E.1 Dataset Configuration Details

Missing Modality Scenario Construction. We use two evaluation settings. The first uses the fully paired subset, where samples contain complete modality triples and missing modalities are simulated during training and evaluation. This setting allows controlled comparison across missing patterns, since every held-out modality has a corresponding observed target representation. The second uses the partially paired cohort, where modalities are absent according to the observed clinical data collection and linkage pattern rather than synthetic masking.

Table 17: Online support and batch-size stability on MIMIC mortality at 50% missingness, equal-weight mean over five seeds and three modalities.

Batch	Raw support	Eff. support	Insuff.-support rate	M50 AUROC	M50 AUPRC
4	2.0	2.0	59.4%	78.4 ±1.9	48.0 ±3.9
8	4.3	4.0	7.6%	80.7 ±0.8	50.0 ±1.9
16	9.0	8.2	0.1%	81.1 ±0.6	50.6 ±2.0
32	17.9	16.5	0.0%	79.1 ±1.4	48.8 ±2.8
64	35.6	33.0	0.0%	78.0 ±2.4	46.2 ±4.9

Table 18: Controlled CF-target variance as a function of the reference-set size R.

Reference count R	CF-target variance	Effective support
2	5.445×10^{-3}	1.93
4	2.664×10^{-3}	3.84
8	1.269×10^{-3}	7.65
16	6.526×10^{-4}	15.35
32	3.126×10^{-4}	30.66
64	1.397×10^{-4}	61.24

MIMIC-IV Dataset Configuration. We utilize the MIMIC-IV database [19], a large, publicly available database comprising de-identified health-related data. We use the de-identified MIMIC-IV and MIMIC-CXR datasets under credentialed PhysioNet access and follow the corresponding data use agreements; no new patient data are collected. Our data construction follows MedFuse [3], which inherits the Harutyunyan clinical time-series benchmark [21]. Specifically, the EHR stream uses cohort filtering, hourly discretization, carry-forward imputation, and binary observation masks. CXR studies are linked to ICU stays using patient identifiers and split at the patient level to avoid leakage. We extend the EHR–CXR setting with radiology reports, which are naturally paired one-to-one with CXR studies in MIMIC-CXR. Thus, CXR and TXT are jointly present or jointly absent in our tri-modal setup. We employ three complementary modalities:

- **Structured time-series Electronic Health Records (EHR):** Contains vital signs and laboratory results from the MedFuse cohort. Measurements are discretized hourly with carry-forward imputation and observation masks, and normalized using z-score normalization.
- **Chest X-ray images (CXR):** Provides visual diagnostic information from radiographic imaging. Images are resized to 224×224 pixels and normalized with the SigLIP2 image processor (mean and standard deviation 0.5).
- **Radiology reports (TXT):** The free-text radiology report of the paired chest X-ray study. Discharge summaries and nursing notes are not used. Text is tokenized using a clinical BERT tokenizer with maximum sequence length of 512.

Temporal alignment and leakage screening. For in-hospital mortality and length-of-stay, the prediction time is 48 hours after ICU admission, and all inputs are restricted to information available by this time. EHR uses hourly measurements from `intime` to `intime`$+48$h. Among CXR studies within the admission, we keep only those whose study time falls in $[\texttt{intime}, \texttt{intime} + 48\text{h}]$ and pair the most recent one. For the paired radiology report, we additionally require that it was committed to the database before the cutoff, using the `storetime` field of MIMIC-IV-Note, which rules out reports finalized after the prediction time. Because MIMIC-CXR provides no field corresponding to `storetime` while the same reports appear in MIMIC-IV-Note with a `storetime` field, and no crosswalk between MIMIC-CXR `study_id` and MIMIC-IV-Note `note_id` is released, we link reports by content within each patient and include a study only when the matching procedure identifies a unique MIMIC-IV-Note radiology note.

- **In-Hospital Mortality Prediction:** Binary classification predicting whether a patient survives to hospital discharge. We use 4,885 training, 540 validation, and 1,373 test samples.

Table 19: Mask ratio sensitivity on MIMIC mortality at 50% missing modality ratio. Starred entries denote the default setting used in the main experiments.

Varied Modality	0.1	0.25	0.5	0.75
EHR (CXR=0.75, TXT=0.25)	83.8	83.1	81.1*	81.0
CXR (EHR=0.5, TXT=0.25)	83.3	83.8	83.6	81.1*
TXT (EHR=0.5, CXR=0.75)	83.8	81.1*	81.6	82.9

Table 20: Per-missing-configuration breakdown for CEPA on MIMIC mortality. AUROC is reported as mean±std across five seeds.

Observed modalities	AUROC (↑)
EHR+CXR	81.3 ±1.1
EHR+TXT	84.9 ±1.0
CXR+TXT	84.4 ±2.1
EHR only	76.3 ±0.1
CXR only	75.9 ±2.4
TXT only	82.0 ±1.3

- **Length-of-Stay Prediction:** Regression of the remaining ICU stay after the 48-hour observation window, on the same split as mortality. MAE is reported in hours, and κ is the linearly weighted Cohen's kappa over 10 day-level bins.
- **Phenotyping:** Multi-label classification simultaneously predicting 25 clinical conditions (e.g., acute renal failure, chronic kidney disease, respiratory failure). We use 7,744 training, 882 validation, and 2,166 test samples.

E.2 Missing Modality Simulation and Evaluation Protocol

To ensure fair comparison across methods, we first extract samples for which all three modalities are available, and then simulate missing modality scenarios during training. Specifically, each training batch is composed of equal proportions (1/3 each) of three conditions: (i) all modalities present, (ii) exactly one modality missing, and (iii) exactly two modalities missing. Within each condition, the specific missing combination is sampled uniformly at random. This balanced protocol forces the model to learn generalizable cross-modal relationships across all levels of data incompleteness rather than memorizing specific missing configurations.

For evaluation, we design systematic missing scenarios that reflect real-world deployment challenges. At missing ratio r, a fraction r of the test samples has missing modalities, and the remaining samples keep all three modalities. The main results use $r = 50\%$, and Figure 4 varies r from 0% to 90%. The samples with missing modalities are split equally between two conditions:

- **Partial Missing Scenarios:** Exactly one modality is missing (half of the samples with missing modalities), simulating common situations like equipment failure, data corruption, or acquisition constraints.
- **Severe Missing Scenarios:** Exactly two modalities are missing (the other half), representing critical situations where only minimal information is available.

This balanced protocol ensures comprehensive assessment across different levels of data incompleteness and provides insights into model degradation patterns under increasing data scarcity. We report the mean and standard deviation over five seeds (paired significance tests in Appendix D.4).

E.3 Baseline Implementation Details

All baselines use our shared encoder backbone (Transformer for EHR, SigLIP2 for CXR, CXR-BERT for TXT) with a common 256-dimensional feature space; Table 21 summarizes their adaptations.

Table 21: Summary of baseline methods and adaptations to the trimodal heterogeneous setting.

Method	Strategy	Core Mechanism	Key Adaptation
SMIL [6]	Meta-learning	Bayesian prototype imputation	Learnable prototype banks; bi-level opt. for $\geq$2 modalities
MUSE [11]	Graph contrastive	Patient–modality bipartite graph + InfoNCE	EdgeSAGE on projected features; excluded from LoS[†]
ShaSpec [9]	Disentanglement	Shared/specific encoders + domain classifier	Shared MLP replaces 3D conv; proxy feature for missing mod.
DRIM [10]	Disentanglement	Adversarial shared/unique decomposition	Deep encoder copies; two-scale masked attention preserved
M3Care [4]	Imputation	Task-guided deep kernels + GCN	Common-dim projection before similarity computation
DiCMoR [22]	Generative	Glow normalizing flows + class-cond. prior	Per-modality flows with shared prior; excluded from LoS[†]
FuseMoE [23]	MoE	Sparse MoE + Laplace gating	Learnable tokens for missing mod. routed through experts
Flex-MoE [24]	MoE	Missing modality bank + sparse expert routing	Tri-modal bank indexed by observed/missing mod. combination
SimMLM [7]	MoE	Dynamic MoE + MoFe ranking loss	Heterogeneous encoders replace U-Net experts
Robult [25]	PID	Redundancy/Unique/Synergy decomposition	Attention fusion with residual MLP; fully supervised mode
Cross-modal representation learning (not originally designed for missing modalities)			
M3-JEPA [27]	JEPA	Joint embedding prediction + MMoE predictor	All 6 pairwise directions; alternating gradient descent

[†]Excluded from length-of-stay prediction due to requirement of discrete class labels.

Non-trivial adaptations. Several baselines required substantial modifications beyond encoder replacement:

- **ShaSpec**: The original shared 3D convolutional encoder is replaced with a shared MLP over heterogeneous encoder outputs. The compositional layer fuses shared and specific features via $\hat{h}_m = h_m^s + \text{MLP}([h_m^s \| h_m^d])$; missing modalities use the first available modality's shared feature as proxy.
- **M3-JEPA**: Extended from bimodal to all 6 pairwise prediction directions, computing JEPA loss only when both source and target modalities are available.
- **DiCMoR**: Per-modality Glow flows share a single class-conditional prior, with cross-modal Transformer fusion applied after reconstruction.
- **Flex-MoE**: We adapt the missing modality bank to EHR, CXR, and TXT combinations and keep the sparse expert routing design.

E.4 Synthetic Experiment Details

Data generation. We sample a 2D latent variable $\mathbf{s} = (s_1, s_2)$ on a ring with radius $r \sim \mathcal{N}(1.0, 0.04)$ and uniformly random angle $\theta \in [0, 2\pi)$. The angle determines the class label among C=16 equally-spaced sectors. Each of the 12 modalities observes $\mathbf{s}$ through a 1D linear projection $z_m = \mathbf{w}_m^\top \mathbf{s} + \epsilon_m$, where $\mathbf{w}_m = (\cos\frac{m\pi}{12}, \sin\frac{m\pi}{12})$ and $\epsilon_m \sim \mathcal{N}(0, 0.0225)$. Observations are quantized (q=0.05) and clipped to $[-2, 2]$. We generate 5,000 training and 2,000 test samples.

Model architecture. Each scalar observation z_m is embedded into a 32-dimensional vector via a modality-specific linear projection (implemented as grouped 1D convolution). A mean-pooling context aggregator computes a context vector $\mathbf{c} \in \mathbb{R}^{64}$ from the observed embeddings. A two-layer MLP predictor maps $\mathbf{c}$ to predicted embeddings for all 12 modalities. The filled embeddings (observed + imputed) are processed by a 2-layer Transformer fusion module with 4 heads, followed by a linear classifier.

Fill strategies. We compare seven strategies: (1) *Zero*: fill with zeros; (2) *Mean*: fill with per-modality mean embedding computed from training data; (3) *Learned*: fill with a learnable token per modality; (4) *MSE*: use the predictor output with supervised MSE loss against ground-truth embeddings; (5) *EP*: marginal characteristic function matching via random projections; (6) *MMD*: marginal moment matching using a Gaussian kernel between predicted and observed embeddings; (7) *CF* (Ours): context-conditional CF matching as described in Section 2.4. All methods share the same backbone architecture and differ only in the fill mechanism and auxiliary loss.

Training. All models are trained for 100 epochs with AdamW (lr=1e-3, batch size 128). During training, each modality is independently dropped with probability p_{miss}=0.5, ensuring at least one modality is observed. The reconstruction loss weight is 1.0. For evaluation, we fix exactly M observed modalities (randomly chosen) and report accuracy averaged over 10 random mask trials.

F Implementation Details

F.1 Model Architecture Details

Modality-Specific Encoders. For MIMIC-IV, we employ domain-specific encoders optimized for each data type:

- **EHR Encoder:** A 2-layer Transformer encoder with 4 attention heads and hidden dimension 256. We use temporal positional encoding to capture patterns in vital signs and lab results.
- **CXR Encoder:** SigLIP2 [28] vision transformer initialized from weights pretrained on large-scale image-text pairs, with input resolution 224×224.
- **TXT Encoder:** CXR-BERT [29], using the `CXR-BERT-Specialized` checkpoint trained on clinical text, with maximum sequence length 512 and hidden dimension 768.

The TXT encoder is initialized from public pretrained weights, and the EHR encoder is trained from scratch. For each MIMIC task, the CXR encoder of CEPA and all baselines is initialized from the corresponding CXR-only model, trained from the public SigLIP2 weights on the training partition of the same patient-level split, and the same checkpoint is used for all five seeds. After this initialization, all components are trained end-to-end.

Cross-Modal Components. The representation predictor g_ϕ is implemented as a 2-layer Transformer with:

- Hidden dimension: 256
- Attention heads: 8
- Dropout rate: 0.3
- Layer normalization applied before each sub-layer
- Residual connections around each sub-layer

The mask predictor h_ω is a scoring head on the token outputs of the EMA predictor $g_{\bar{\phi}}$ that produces one masking score per token. The fusion module f_ψ is a single-layer Transformer that aggregates multimodal representations for final prediction, with output dimension matching the number of classes for each task.

F.2 Training Configuration Details

Optimization Settings. We train all models using the Adam optimizer [44] with the following task-specific configurations:

Table 22: Task-specific training configurations.

Task	Learning Rate	Batch Size	Epochs
MIMIC-IV Mortality	1×10^{-5}	16	100
MIMIC-IV Phenotyping	5×10^{-5}	16	100
MIMIC-IV Length-of-Stay	5×10^{-5}	16	100

We apply gradient clipping with maximum norm 1.0 to prevent gradient explosion. Early stopping is employed based on validation performance with patience of 10 epochs. Experiments were conducted on NVIDIA RTX 3090 24GB and NVIDIA RTX A6000 48GB GPUs, with training time varying by dataset, modality configuration, and missingness setting. Inference cost and training overhead are reported in Appendix D.2.

Regularization. We apply dropout with rate 0.3 to all Transformer layers. For the loss weights, we set $\alpha = 0.01$ for observed modality reconstruction ($\mathcal{L}_{\text{obs}}$), $\beta = 0.01$ for cross-modal consistency ($\mathcal{L}_{\text{cross}}$), and $\gamma = 0.1$ for the context-conditional CF alignment ($\mathcal{L}_{\text{CF}}$). These weights were determined through grid search on validation sets.

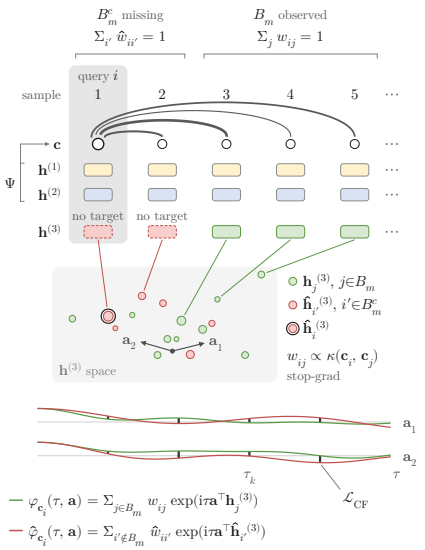


Figure 8: **Context-conditional CF matching on a mini-batch.** Columns are samples that miss modality $m{=}3$ (B_m^c, including the query i) or observe it (B_m), and rows are the context $\mathbf{c}$, built by Ψ from the observed embeddings, and the pooled embeddings. The imputations $\widehat{\mathbf{h}}^{(3)}$ have no target. The NW weights w_{ij} and $\widehat{w}_{ii'}$, drawn as arc width and point radius, are normalized within B_m and B_m^c, respectively, and detached from the computation graph. $\mathcal{L}_{\text{CF}}$ averages the squared gaps between the two CF estimates (real parts shown) over projections $\mathbf{a}_\ell$, frequencies τ_k, and queries.

The mask predictor uses exponential moving average (EMA) updates with decay coefficient $\lambda = 0.996$ to provide stable masking guidance. For $\mathcal{L}_{\text{cross}}$, we mask a larger fraction of tokens than for $\mathcal{L}_{\text{obs}}$. This dual-ratio strategy encourages the model to rely more heavily on predicted missing representations during cross-modal consistency learning.

F.3 Implementation of Context-Aware Conditional Distribution Alignment

Computational considerations. The proposed context-conditional CF alignment involves computing pairwise context similarities within a mini-batch, which incurs a quadratic complexity in the batch size. Specifically, evaluating the kernel weights $\kappa(\mathbf{c}_i, \mathbf{c}_j)$ for all sample pairs scales as $\mathcal{O}(n^2)$ for a minibatch of $n = |\mathcal{B}|$ samples. In practice, this cost can be mitigated by restricting the kernel computation to a subset of nearest neighbors in the context space or by subsampling reference samples j when estimating the conditional ECF, without affecting the overall formulation.

Stability and implementation details. As in Section 2.4, the kernel weights are normalized for each query and detached from the computation graph. We use $N_a = 128$ random projection directions sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$, $N_\tau = 64$ frequencies τ_k drawn uniformly from $[-\tau_{\max}, \tau_{\max}]$ with $\tau_{\max} = 5.0$, and kernel bandwidth $\sigma = 1.0$ for the Gaussian context similarity kernel. Both projections and frequencies are resampled at each training step to ensure diverse coverage of the distribution space.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of the method, the missing-modality setting, and the main empirical claims. The conclusion and Appendix D.12 further qualify the scope through limitations and broader impacts.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the conclusion and Appendix D.12, including task-oriented representation prediction versus faithful modality recovery, the evaluation scope (MIMIC and CMU-MOSEI), and the minibatch support required by the conditional CF objective, whose pairwise cost is discussed in Appendix F.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the theoretical result on gradient boundedness in Section 2.4, and the full proof is provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The datasets, baselines, missing-modality protocol, model architecture, and training configuration are described in the experiments section and Appendices E and F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at https://github.com/jungwon-choi/CEPA. The MIMIC datasets are available through credentialed PhysioNet access under the corresponding data use agreements.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies dataset splits, missing-modality simulation, architectures, optimizer, hyperparameters, and baseline adaptations in the experiments section and appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported as mean and standard deviation across five independent runs with different random seeds. Seed-paired tests and 95% confidence intervals for the main comparisons are reported in Appendix D.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F reports the GPU resources used and includes throughput, parameter overhead, inference speed, and training wall-clock overhead analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses de-identified, credentialed-access clinical datasets, follows the corresponding data use agreements, and does not collect new patient data. We have considered the NeurIPS Code of Ethics and do not identify any deviations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix D.12 discusses potential benefits for robust clinical prediction under missing data and risks related to standalone clinical use, dataset bias, distribution shift, and the need for validation and governance.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release high-risk generative models, scraped datasets, or other assets with a high risk of misuse. The clinical datasets used are existing de-identified datasets accessed under credentialed data use agreements.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing datasets, pretrained encoders, and baseline methods are cited in the paper and used under their corresponding terms of access or use. The MIMIC-IV,

MIMIC-CXR, and MIMIC-IV-Note datasets are used under the PhysioNet Credentialed Health Data License 1.5.0 and the corresponding data use agreements. The SigLIP2 and CXR-BERT checkpoints are released under the Apache 2.0 and MIT licenses, respectively, and CMU-MOSEI is accessed through the CMU Multimodal SDK (MIT License) with pre-extracted features released by GCNet.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Code is released at `https://github.com/jungwon-choi/CEPA`, and the paper describes the model architecture, training configuration, datasets, and limitations needed to document its use. No new dataset is introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing experiments or newly recruited human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work uses existing de-identified clinical datasets available under credentialed access and does not involve new data collection, intervention, or interaction with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used for grammar, spelling, and word-choice editing, drafting assistance for presentation, and assistance in improving the clarity of proof descriptions. The core method, theoretical claims, experiments, and final technical content were developed and verified by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.