PREVENTING REPRESENTATION COLLAPSE IN LATENT PREDICTION VIA CONTEXT-CONDITIONAL ALIGNMENT UNDER MISSING MODALITIES

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

ABSTRACT

Multimodal models deployed in real-world settings 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 models struggle when modality-specific information is absent. A key challenge is that, 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 via adaptive generative imputation directly in latent space. CEPA employs masked representation learning with data-driven masking patterns and a context-conditional distribution alignment objective that stabilizes latent prediction and reconstructability of observed modalities. Experiments on the heterogeneous MIMIC benchmark—spanning EHR time-series, chest X-rays, and clinical notes—show that CEPA consistently outperforms prior methods across four prediction tasks, with ablations confirming the importance of adaptive masking and alignment.

1 INTRODUCTION

Multimodal learning has emerged as a powerful paradigm for integrating complementary information from heterogeneous sources, such as images, text, and time-series signals (Wu et al., 2024a; Zong et al., 2024). By exploiting cross-modal interactions, multimodal approaches yield richer representations and superior predictive performance compared to unimodal methods (Wu et al., 2024a). Yet, most existing frameworks rely on the strong assumption that all modalities are simultaneously available during both training and inference. In practice, however, this assumption is rarely satisfied. Modalities may be absent due to acquisition costs, privacy restrictions, sensor malfunction, or data corruption (Hayat et al., 2022; Zhang et al., 2022; Yao et al., 2024). For instance, many patients' medical records lack imaging scans, while online media data may include text but not audio. Such incomplete multimodal conditions often lead to severe performance degradation, undermining the reliability of multimodal systems in real-world applications (Ma et al., 2021; Li et al., 2025).

Existing approaches to missing modality scenarios can be categorized into three main paradigms. *Generative imputation methods* (Boyko et al., 2025; Yao et al., 2024) reconstruct missing raw data through autoencoder architectures, but they may be expensive for high-dimensional modalities and exhibit semantic drift in generated content that degrades downstream task performance. *Feature alignment approaches* (Wang et al., 2023a; Zhang et al., 2022) project modalities into shared representation spaces using linear or nonlinear transformations, assuming modalities contain overlapping semantic information—an assumption that fails for modalities with disjoint information content or different temporal resolutions. *Domain-specific architectures* (Hayat et al., 2022; Xu et al., 2024; Li et al., 2025) incorporate task-specific inductive biases through specialized attention mechanisms or fusion layers, limiting their applicability to new domains without architectural modifications.

While the three paradigms above cover the dominant directions, existing methods also adopt related strategies that attempt to refine or extend them. Meta-learning approaches (Ma et al., 2021; Zhang et al., 2022) require extensive support sets with complete modality pairs during training, limiting

their applicability to scenarios with sparse supervision. Disentanglement-based methods (Wang et al., 2023a; Yao et al., 2024; Xu et al., 2024; Li et al., 2025) optimize modality-specific reconstruction objectives that fail to generalize across heterogeneous data distributions. Masked autoencoder variants (Boyko et al., 2025; Wu et al., 2024b) achieve reconstruction fidelity through computationally expensive generative modeling in high-dimensional input spaces, while contrastive methods depend critically on carefully designed positive-negative pair sampling strategies. These approaches fundamentally treat missing modalities as a data corruption problem rather than leveraging modality incompleteness as a regularization mechanism for learning robust cross-modal representations. Across these approaches, missing-modality learning is typically framed as a reconstruction or corruption problem. However, when a modality is unobserved, there is no uniquely identifiable ground-truth representation to predict. As a result, pointwise reconstruction objectives are inherently ill-posed and often induce representation collapse.

In this work, we introduce CEPA, **C**ross-modal **E**mbedding **P**rediction and **A**lignment, a robust learning framework designed to handle multimodal data with missing modalities. Inspired by masked representation learning frameworks (He et al., 2022; Assran et al., 2023), our framework generatively imputes the representations of missing modalities from masked inputs. However, unlike conventional masked representation learning that relies on randomly sampled masks, we propose an *adaptive masking strategy* that learns informative masking patterns directly from data, thereby improving the efficiency of multimodal representation learning. To further enhance cross-modal robustness, we optimize a distribution alignment objective that explicitly emphasizes consistency across modalities. Although our method can be viewed as an imputation approach, it differs fundamentally from prior imputation methods that operate in raw input space: instead, we directly impute missing modality representations in the latent space, allowing the model to focus on capturing and recovering core semantic content rather than superficial details. We evaluate our method on the multimodal MIMIC benchmark (integrating EHR, CXR, and clinical notes) across four distinct clinical tasks, demonstrating consistent gains in both resilience and predictive accuracy across a range of missing-modality scenarios.

2 METHODS

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.

Our primary goal is to learn a multimodal learning framework that is robust to missing modalities. Specifically, we aim to learn a *representation predictor* that infers the representations of missing modalities by exploiting cross-modal correlations, approximating the representations that would have been computed from plausible imputations of the missing observations.

In this framework, 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. Accordingly, each modality is treated as a sequence and processed by its corresponding encoder $\mathcal{E}^{(m)}$, yielding a representation $\mathbf{Z}^{(m)} = [\mathbf{z}_1, \ldots, \mathbf{z}_{L^{(m)}}] \in \mathbb{R}^{L^{(m)} \times d_m}$ where d_m denotes the dimensionality of the token-level features for the modality m. Throughout the paper, we refer to $\mathbf{Z}^{(m)}$ as a single (structured) representation, and do not explicitly distinguish between the representation and its constituent token-level elements unless necessary.

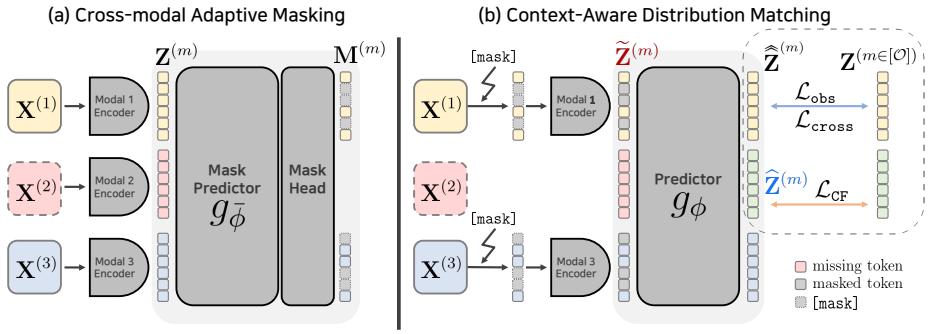

Figure 1: **Overview of CEPA.** Cross-modal embedding prediction architecture provides auxiliary supervision for observed modalities while guiding the representation predictor to impute coherent missing-modality embeddings, ultimately supporting robust multimodal fusion.

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. An overview of the framework is shown in Figure 1.

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 (He et al., 2022). 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 (Assran et al., 2023; Bardes et al., 2025; Assran et al., 2025).

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,

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

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

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,

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

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 $\hat{\mathbf{Z}}^{(m)}$ 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 based on information from 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 when the dataset contains a high proportion of missing data.

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

In principle, one could predict the missing representations directly without masking the observed modalities. Since the representation predictor g_ϕ already provides a mechanism for imputing missing representations, it could be trained to minimize reconstruction error against the ground truth. However, aside from the fundamental issue that no ground-truth representations exist for missing modalities, this approach may also lead to suboptimal solutions.

Specifically, the representations that support task performance and those that are optimal for missing-modality prediction may not coincide, potentially resulting in representation collapse or the learning of task-irrelevant features (Balestriero & LeCun, 2024). To mitigate these issues and improve the robustness of representation prediction, we introduce a cross-modal embedding prediction architecture framework. In this framework, we apply masks to inputs from the observed modalities and train g_ϕ to predict the representations for both observed and missing modalities.

The central idea is that randomly masking portions of the observed modalities and requiring the model to reconstruct them produces a stronger and more stable training signal. This strategy offers several benefits: (1) it discourages the predictor from overfitting to trivial solutions, (2) it promotes the learning of semantically meaningful cross-modal relationships, and (3) it provides additional supervised signals that complement the missing-modality prediction objective.

Nevertheless, naive random masking may still be suboptimal for cross-modal learning. Different modalities vary in information density and semantic relevance. To address this, we propose to learn which portions of the observed modalities should be masked so as to maximally benefit representation prediction for the missing modalities.

2.3.2 Modal-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, $\bar{\phi} \leftarrow \tau\bar{\phi} + (1-\tau)\mathsf{stopgrad}(\phi)$, where $\tau \in [0, 1)$ is the decay coefficient and $\mathsf{stopgrad}(\cdot)$ is the stop gradient operation blocking the gradient flow. Using $\bar{\phi}$, for each observed modality $m \in \mathcal{O}$, we compute

$$\begin{aligned}
\boldsymbol{\sigma}^{(m)} &= h_\omega\Big(g_{\bar{\phi}}\big(\texttt{concat}\big(\{\mathbf{Z}^{(m)}\}_{m\in\mathcal{O}}, \{\mathbf{T}^{(m)}\}_{m\in\mathcal{M}}\big)\big)\Big),\\
\mathcal{I}^{(m)} &= \texttt{TopK}(\boldsymbol{\sigma}^{(m)}, K^{(m)}) \subseteq \{1, \dots, L^{(m)}\},\\
\mathbf{m}_t^{(m)} &= 1[t \in \mathcal{I}^{(m)}],\ \mathbf{M}^{(m)} = [\mathbf{m}_1^{(m)}, \dots, \mathbf{m}_{L^{(m)}}^{(m)}],
\end{aligned} \tag{3}$$

where $\texttt{TopK}(\cdot, K^{(m)})$ 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 $\texttt{TopK}$ selection and indicator function are non-differentiable, we employ a straight-through estimator (STE) (Bengio et al., 2013) to enable gradient flow: the forward pass uses the hard binary mask, while the backward pass propagates gradients through the continuous activation scores $\boldsymbol{\sigma}^{(m)}$ 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 passed through the encoder to obtain $\tilde{\mathbf{Z}}^{(m)}$ (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 $g\phi$ 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 training 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 use *higher* masking ratio (i.e., smaller $K^{(m)}$) than the one used 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 observed 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 let $B_m \subseteq [N]$ denote the index set of samples 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

$$\sum_{j\in B_m} \underbrace{\frac{\kappa(\mathbf{c}_i, \mathbf{c}_j)}{\sum_{j'\in B_m} \kappa(\mathbf{c}_i, \mathbf{c}_{j'})}}_{:=w_{ij}^{(m)}} f(\mathbf{h}_j^{(m)}), \tag{9}$$

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 $\tau > 0$ 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 $B_m^c = [N] \setminus B_m$ 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:

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

where $w_{ii'}^{(m)}$ 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 (Balestriero & LeCun, 2025):

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

Here we average discrepancies over L random projections $\{\mathbf{a}_\ell\}_{\ell=1}^L$ and K sampled frequencies $\{\tau_k\}_{k=1}^K$. The total CF regularization over all missing modalities is then given by $\mathcal{L}_{\mathrm{CF}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathcal{L}_{\mathrm{CF}}^{(m)}$. As highlighted by Balestriero & LeCun (2025), CF-based matching can offer improved geometric stability compared to density-based divergences; accordingly, 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 strictly bounded by the frequency bandwidth. 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. This property ensures a smooth optimization landscape free from gradient explosion (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-specific loss $\mathcal{L}_{\text{task}}$ employs cross-entropy for classification tasks and mean squared error for regression tasks.

Overall objective. The proposed context-aware 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 controls the strength of the 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 and analyze the properties of missing modality prediction strategies, we design a controlled synthetic experiment with a ring-structured latent space (Figure 2).

Setup. A 2D latent variable $\mathbf{s}$ lies on a ring, with its angle determining the class label among C=16 classes. Each of M=12 modalities observes $\mathbf{s}$ through a distinct linear projection $z_m = \mathbf{w}_m^\top \mathbf{s} + \epsilon_m$, with additive noise and quantization. Each modality is independently embedded, a context vector is computed from observed embeddings via set encoder, and a predictor network maps this context to the missing embeddings. Classification accuracy is measured by varying the number of observed modalities k.

Results. We compare six imputation strategies—Zero, Mean, Learned token, MSE (supervised reconstruction with ground-truth targets), EP (marginal characteristic function matching), and CF (our context-conditional CF matching)—in Figure 2(c). CF produces the most coherent decision boundaries closest to the ground truth (GT), particularly at low k. Static fill strategies (Zero, Mean, Learned) yield fragmented boundaries, and even supervised MSE falls short of CF because it does not enforce distributional consistency. Figure 2(b) further shows that CF matching *without* stop-gradient on the context leads to representation collapse (context diversity $\to$ 0), whereas applying stop-gradient (as in our framework) preserves context diversity throughout training, consistent with the analysis in Section 2.4. Full experimental details are provided in Appendix C.4.

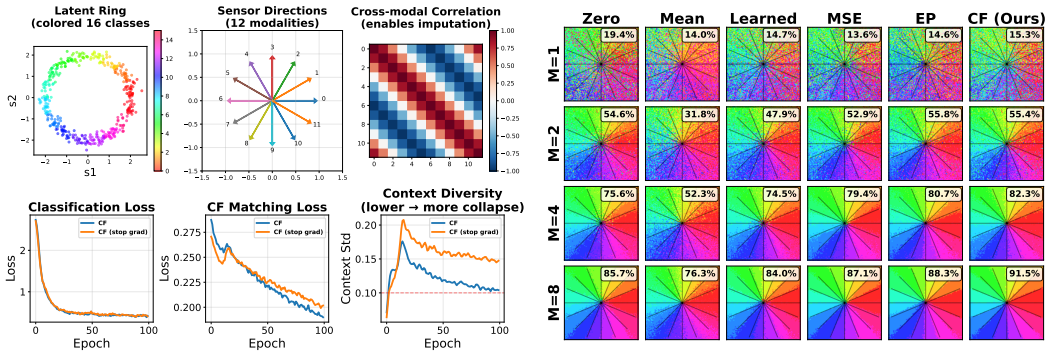

Figure 2: **Synthetic experiment. (a)** Data generation: latent ring structure, 12 sensor directions, and cross-modal correlation matrix. **(b)** Context collapse analysis: CF matching without stop-gradient leads to representation collapse (context diversity $\to 0$), while stop-gradient preserves it. **(c)** Decision boundaries across fill strategies and number of observed modalities k. CF (Ours) produces the most coherent boundaries.

3.2 Datasets and Experimental Setup

Multimdal MIMIC Dataset and Tasks. We utilize the MIMIC-IV and MIMIC-CXR database (Johnson et al., 2023; 2024) with three complementary modalities: structured time-series EHR, chest X-ray images (CXR), and clinical text reports (TXT). Following Hayat et al. (2022), 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). These tasks span classification and regression, demonstrating our framework's generalizability. Dataset statistics and detailed configurations are provided in Appendix C.

3.3 Baselines

We benchmark CEPA against two groups of baselines. The first group comprises 9 *missing modality methods* that explicitly handle modality incompleteness: *meta-learning* (SMIL (Ma et al., 2021)), *graph-based contrastive learning* (MUSE (Wu et al., 2024b)), *shared-specific disentanglement* (ShaSpec (Wang et al., 2023a), DRIM (Robinet et al., 2024)), *patient similarity-based imputation* (M3Care (Zhang et al., 2022)), *generative recovery* (DiCMoR (Wang et al., 2023b)), *mixture-of-experts* (FuseMoE (Han et al., 2024), SimMLM (Li et al., 2025)), and *partial information decomposition* (Robult (Nguyen et al., 2025)). The second group includes *cross-modal representation learning* methods not originally designed for missing modalities: M3-JEPA (Lei et al., 2025), which we adapt to our missing modality setting for a broader comparison. Methods requiring discrete labels (MUSE, DiCMoR) are excluded from length-of-stay prediction. We refer readers to Appendix 4 for a comprehensive discussion of related work and to Appendix C.3 for detailed adaptation and hyperparameter choices for each baseline.

3.4 Implementation Details

Model Architecture. For MIMIC-IV, we employ domain-specific encoders: a Transformer encoder for structured EHR time-series data, SigLIP2 (Tschannen et al., 2025) for chest X-ray images, and CXR-BERT (Boecking et al., 2022) for clinical text 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 without a separate pretraining stage. The representation predictor g_ϕ and mask predictor h_ω are implemented as multi-layer Transformers with hidden dimension 256. The fusion module f_ψ is a single-layer Transformer that aggregates multimodal representations for final prediction.

Training Configuration. We train all models using the Adam optimizer with batch size 16. Learning rates are task-specific: 1×10^{-5} for mortality prediction and 5×10^{-5} for phenotyping. Models are trained for 100 epochs with early stopping based on validation performance. We apply dropout (rate=0.3) and set loss weights $\alpha = 0.01$, $\beta = 0.01$, and $\gamma = 0.1$ for $\mathcal{L}_{\text{obs}}$, $\mathcal{L}_{\text{cross}}$, and $\mathcal{L}_{\text{CF}}$ respectively. For masking configuration, we apply different masking ratios for $\mathcal{L}_{\text{obs}}$ and $\mathcal{L}_{\text{cross}}$, with EMA decay τ=0.996 for the mask predictor (see Appendix D).

Table 1: Performance comparison across three clinical tasks at 50% missing modality ratio on Multimodal MIMIC Benchmark. We report mean±std across three 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
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
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
✔	✔	✔	**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
Cross-modal Masking (Ours)	**81.1**±0.6	**50.6**±2.0

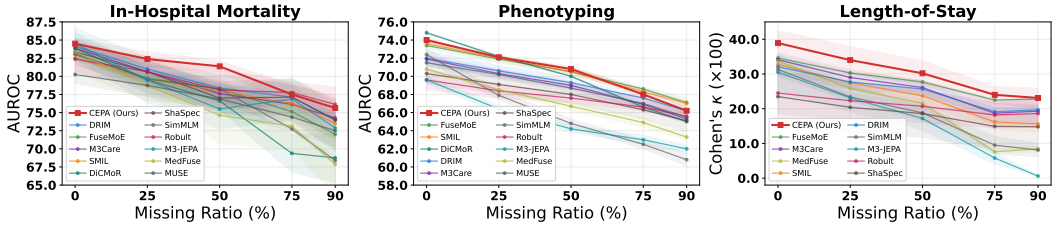

Figure 3: Performance degradation across increasing missing ratios (0%–90%) on clinical prediction tasks of multimodal MIMIC Dataset. CEPA maintains robust performance even under severe missing conditions.

Missing Modality Protocol. To ensure fair comparison, we extract samples with all three modalities present and 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, we balance partial missing (one modality absent, 50% of test samples) and severe missing (two modalities absent, 50%) scenarios. All results are averaged over three runs with different seeds. Full protocol details are in Appendix C.

3.5 Clinical Prediction Results on Multimodal MIMIC Data

Table 1 presents comprehensive comparisons across all three clinical tasks under 50% missing modality ratio, where our method consistently outperforms state-of-the-art baselines. Performance degradation analysis across varying missing ratios (0%–90%) is provided in Figure 3.

3.6 Analysis

Main results. Table 1 shows that CEPA consistently outperforms all baselines across every task and metric. Among existing methods, those that explicitly model cross-modal relationships (e.g., DRIM, FuseMoE) generally perform better, yet none enforce distributional consistency for predicted missing representations. CEPA addresses this gap through context-conditional CF matching ($\mathcal{L}_{\text{CF}}$), yielding consistent gains across both classification and regression tasks.

Robustness under severe missing ratios. Figure 3 shows that CEPA maintains strong performance even when the missing ratio increases to 90%, where two out of three modalities are absent for most test samples. While all baselines exhibit substantial degradation under such severe conditions, our method retains competitive accuracy thanks to the context-conditional CF alignment, which provides a distributional prior for the predicted representations regardless of how many modalities are observed.

Masking ablation. Table 3 compares masking strategies. Without masking, the predictor can trivially copy observed representations without learning cross-modal relationships. Random and block masking progressively improve performance, but our cross-modal masking—which leverages information from other modalities to select informative tokens to mask—yields the largest gain (+1.9 AUROC, +5.3 AUPRC over no masking), confirming that data-driven, cross-modal-aware masks force the predictor to capture more meaningful inter-modal dependencies.

Loss ablation. Table 2 disentangles the contributions of the three loss components. Starting from $\mathcal{L}_{\text{obs}}$ alone, adding $\mathcal{L}_{\text{cross}}$ improves both metrics by encouraging the predictor to reconstruct missing representations from observed ones. Replacing $\mathcal{L}_{\text{cross}}$ with $\mathcal{L}_{\text{CF}}$ yields a larger gain, demonstrating that distributional alignment via characteristic function matching provides a stronger supervisory signal than pointwise cross-modal reconstruction. The full model combining all three objectives achieves the best performance, confirming that $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{CF}}$ play complementary roles: the former enforces sample-level semantic coherence while the latter aligns the distribution of predicted representations to the empirical conditional distribution of observed ones.

4 Related Work

4.1 Modeling under Missing Modalities

Missing modalities remain a key challenge in deploying multimodal systems across clinical, affective, and perceptual tasks (Yao et al., 2024; Zhang et al., 2022). Early work used variational or adversarial generators to recover missing views (Dorent et al., 2019; Sharma & Hamarneh, 2019), but pixel- or token-level synthesis is noisy and computationally heavy (Yao et al., 2024). Recent methods instead reason in latent space through diverse strategies. Disentanglement-based approaches decompose representations into shared and modality-specific factors (Wang et al., 2023a; Yao et al., 2024; Robinet et al., 2024). Knowledge distillation methods transfer knowledge from complete to incomplete modality settings via Bayesian meta-learning or cross-modal attention alignment (Ma et al., 2021; Zhuang et al., 2025). Graph-based methods model inter-modality relations explicitly through modality-adaptive graphs or contrastive losses (Zhang et al., 2022; Wu et al., 2024b). Generative approaches learn flexible missing distributions using conditional normalizing flows (Wang et al., 2023b). Mixture-of-experts approaches dynamically combine unimodal predictors with soft routing and modality gating (Xu et al., 2024; Han et al., 2024; Li et al., 2025).

4.2 Representation Learning for Multimodal Robustness

Self-supervised learning (SSL) leverages unlabeled data to capture modality-agnostic structure, complementing supervised adaptation. Contrastive objectives align partial and complete views of the same instance to prevent modality collapse (Wu et al., 2024b; Lin & Hu, 2023), while other works design augmentations that keep embeddings stable under structured dropout (Nguyen et al., 2025). Masked autoencoding treats missing modalities as extreme masks: impuTMAE (Boyko et al., 2025) extends ViT-MAE to heterogeneous medical inputs, and analyses show masking noise emphasizes perceptually informative subspaces (Balestriero & LeCun, 2024). Joint-Embedding Predictive Architectures (JEPA) instead predict latent codes of missing views from available ones, avoiding contrastive negatives and remaining effective with partial observations (Lei et al., 2025).

5 Conclusion

We presented CEPA, a multimodal learning framework enhanced with adaptive masking and representation imputation to handle missing-modality scenarios. Our approach learns masking patterns tailored to observed modalities while imputing missing representations through cross-modal interactions. The key contributions include multi-objective representation prediction with complementary losses, cross-modal consistency regularization that enables training without ground-truth for missing modalities. Experiments on MIMIC-IV demonstrate consistent gains over state-of-the-art baselines under diverse missing-modality conditions.

REFERENCES

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*, pp. 15619–15629, 2023.

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.

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

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

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.

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

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

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.

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*, pp. 74–82. Springer, 2019.

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.

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*, pp. 479–503. PMLR, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q, 2024.

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.

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.

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.

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*, pp. 24068–24077, 2025.

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.

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

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*, pp. 5985–5993, 2025.

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*, pp. 163–173. Springer, 2024.

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.

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.

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*, pp. 15878–15887, 2023a.

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*, pp. 22025–22034, 2023b.

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

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*, 2024b.

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*, pp. 438–446, 2024.

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

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*, pp. 2418–2428, 2022.

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*, pp. 4626–4636, 2025.

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.

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 strictly bounded by the frequency bandwidth. 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 the boundedness 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 $B_m^c := [N] \setminus B_m$:

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

where $w_{ii'}^{(m)} = \kappa(\mathbf{c}_i, \mathbf{c}_{i'}) / \sum_{i''} \kappa(\mathbf{c}_i, \mathbf{c}_{i''})$ 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 $s_i := \mathbf{a}^\top \widehat{\mathbf{h}}_i^{(m)}$ and $\Delta := \widehat{\varphi}_{\mathbf{c}_i}^{(m)} - \varphi_{\mathbf{c}_i}^{(m)}$. The gradient of the predicted CF has two potential pathways:

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

We apply stop-gradient to the kernel weights, i.e., $w_{ij}^{(m)} \leftarrow \mathsf{stopgrad}(w_{ij}^{(m)})$. This eliminates the first term since the weights depend on context $\mathbf{c}$, and allowing gradients through this path would enable the model to manipulate context representations to trivially minimize the loss. With stop-gradient:

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

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

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

The kernel weighting introduces an additional factor $w_{ii}^{(m)} \leq 1$, which can only *reduce* the gradient magnitude compared to the point-mass case. The gradient norm is linear in the frequency bandwidth $\max_k |\tau_k|$ and independent of the context distribution or kernel bandwidth, ensuring stable optimization with a smooth loss landscape. □

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 τ
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 all m

  // Cross-modal Masking (Section 2.2)
  $\mathbf{M}^{(m)} \leftarrow h_\omega(g_{\bar{\phi}}(\mathbf{Z}, \mathbf{T}))$ for $m \in \mathcal{O}$ Eq. equation 1
  $\tilde{\mathbf{Z}}^{(m)} \leftarrow \mathcal{E}^{(m)}(\mathbf{X}^{(m)} \odot \bar{\mathbf{M}}^{(m)} + \mathbf{V}^{(m)} \odot \mathbf{M}^{(m)})$

  // Cross-modal Prediction (Section 2.1)
  $\widehat{\mathbf{Z}} \leftarrow g_\phi(\tilde{\mathbf{Z}}^{\mathcal{O}}, \mathbf{T}^{\mathcal{M}})$ Eq. equation 2
  $\mathcal{L}_{\text{obs}} \leftarrow \|(\widehat{\mathbf{Z}}^{(m)} - \mathbf{Z}^{(m)}) \odot \mathbf{M}^{(m)}\|_F^2$ for $m \in \mathcal{O}$

  // Cross-modal Consistency (Section 2.1)
  $\widehat{\widehat{\mathbf{Z}}} \leftarrow g_\phi(\tilde{\mathbf{Z}}^{\mathcal{O}}_{\text{high-mask}}, \widehat{\mathbf{Z}}^{\mathcal{M}})$ Eq. equation 5
  $\mathcal{L}_{\text{cross}} \leftarrow \|(\widehat{\widehat{\mathbf{Z}}}^{(m)} - \mathbf{Z}^{(m)}) \odot \mathbf{M}^{(m)}\|_F^2$ for $m \in \mathcal{O}$

  // CF Alignment & Task (Section 2.3)
  $\mathcal{L}_{\text{CF}} \leftarrow$ CF matching loss Eq. equation 12
  $\hat{y} \leftarrow f_\psi(\mathbf{Z}^{\mathcal{O}}, \widehat{\mathbf{Z}}^{\mathcal{M}})$; $\mathcal{L}_{\text{task}} \leftarrow \text{Loss}(\hat{y}, y)$

  // Update
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{obs}} + \beta\mathcal{L}_{\text{cross}} + \gamma\mathcal{L}_{\text{CF}}$
  Update ϕ, ψ, ω; $\bar{\phi} \leftarrow \tau\bar{\phi} + (1-\tau)\phi$
end for

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 Detailed Experimental Setup

C.1 Dataset Configuration Details

Missing Modality Scenario Construction. To rigorously evaluate multimodal systems under missing modality conditions, we construct controlled experimental scenarios using datasets that contain complete modality pairs. This approach ensures fair comparison across different missing patterns and eliminates confounding factors from naturally occurring missing data. We utilize the MIMIC benchmark, which combines heterogeneous medical data sources (MIMIC-IV for structured EHR, MIMIC-CXR for chest X-rays, and associated clinical notes), to evaluate our framework across diverse task types including binary classification (mortality prediction), multi-label classification (phenotyping), and regression (length-of-stay prediction).

MIMIC-IV Dataset Configuration. We utilize the MIMIC-IV database (Johnson et al., 2023), a large, publicly available database comprising de-identified health-related data associated with over 200,000 critical care patients. Following the same data processing and experimental setup as Hayat et al. (2022), we employ three complementary modalities:

- **Structured time-series Electronic Health Records (EHR):** Contains vital signs, laboratory results, and medication information collected during patient stays. We extract time-series features using sliding windows and normalize values 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 using ImageNet statistics.
- **Clinical text reports (TXT):** Includes discharge summaries and nursing notes that capture clinical reasoning and patient narratives. Text is tokenized using clinical BERT tokenizer with maximum sequence length of 512.

To construct missing modality scenarios, we extract paired samples containing all three modalities from the complete dataset, ensuring that every sample has ground-truth representations for all modalities during training. We evaluate on three clinical prediction tasks:

- **In-Hospital Mortality Prediction:** Binary classification predicting whether a patient will survive their ICU stay. We use 4,880 training, 540 validation, and 1,373 test samples.
- **Length-of-Stay Prediction:** Regression task estimating the remaining duration of a patient's ICU stay after a fixed 48-hour observation window. We use 4,885 training, 540 validation, and 1,373 test samples.
- **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.

C.2 Missing Modality Simulation and Evaluation Protocol

Our experimental design carefully balances training robustness with evaluation comprehensiveness. 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. We construct two primary evaluation conditions:

- **Partial Missing Scenarios:** Exactly one modality is missing (50% of test samples), simulating common situations like equipment failure, data corruption, or acquisition constraints.
- **Severe Missing Scenarios:** Exactly two modalities are missing (remaining 50% of test samples), 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. All experiments are conducted with three independent runs using different random seeds, and we report mean performance and standard deviation across runs to ensure statistical significance and reproducibility.

C.3 Baseline Implementation Details

We adapt state-of-the-art multimodal fusion methods to our heterogeneous trimodal medical setting. All methods use our shared encoder backbone (Transformer for EHR, SigLIP2 for CXR, CXR-BERT for TXT) with features projected to a common 256-dimensional space. Table 4 summarizes each baseline's core mechanism and key adaptation to our setting.

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

Table 4: Summary of baseline methods and adaptations to the trimodal heterogeneous setting.

Method	Strategy	Core Mechanism	Key Adaptation
SMIL (Ma et al., 2021)	Meta-learning	Bayesian prototype imputation	Learnable prototype banks; bi-level opt. for ≥2 modalities
MUSE (Wu et al., 2024b)	Graph contrastive	Patient–modality bipartite graph + InfoNCE	EdgeSAGE on projected features; excluded from LoS†
ShaSpec (Wang et al., 2023a)	Disentanglement	Shared/specific encoders + domain classifier	Shared MLP replaces 3D conv; proxy feature for missing mod.
DRIM (Robinet et al., 2024)	Disentanglement	Adversarial shared/unique decomposition	Deep encoder copies; two-scale masked attention preserved
M3Care (Zhang et al., 2022)	Imputation	Task-guided deep kernels + GCN	Common-dim projection before similarity computation
DiCMoR (Wang et al., 2023b)	Generative	Glow normalizing flows + class-cond. prior	Per-modality flows with shared prior; excluded from LoS†
FuseMoE (Han et al., 2024)	MoE	Sparse MoE + Laplace gating	Learnable tokens for missing mod. routed through experts
SimMLM (Li et al., 2025)	MoE	Dynamic MoE + MoFe ranking loss	Heterogeneous encoders replace U-Net experts
Robult (Nguyen et al., 2025)	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 (Lei et al., 2025)	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.

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.

C.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 M=12 modalities observes $\mathbf{s}$ through a 1D linear projection $z_m = \mathbf{w}_m^\top \mathbf{s} + \epsilon_m$, where $\mathbf{w}_m = (\cos \frac{m\pi}{M}, \sin \frac{m\pi}{M})$ 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 M 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 six 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) *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. Reconstruction loss weight λ=1.0. For evaluation, we fix exactly k observed modalities (randomly chosen) and report accuracy averaged over 10 random mask trials.

D Implementation Details

D.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 apply temporal positional encoding to capture time-series patterns in vital signs and lab results.
- **CXR Encoder:** SigLIP2 (Tschannen et al., 2025) vision transformer initialized from weights pretrained on large-scale image-text pairs, with input resolution 224×224.

- **TXT Encoder:** CXR-BERT (Boecking et al., 2022), using the `CXR-BERT-Specialized` checkpoint trained on clinical text, with maximum sequence length 512 and hidden dimension 768.

The CXR and TXT encoders are initialized from publicly available pretrained weights, while the EHR encoder is trained from scratch. All components are trained end-to-end without a separate pretraining stage.

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 shares the same architecture as g_ϕ but operates on concatenated representations to generate masking scores. 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.

D.2 Training Configuration Details

Optimization Settings. We train all models using the Adam optimizer (Kingma & Ba, 2015) with the following task-specific 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	1×10^{-5}	16	100

Table 5: Task-specific training configurations.

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.

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-aware CF alignment ($\mathcal{L}_{\text{CF}}$). These weights were determined through grid search on validation sets.

The mask predictor uses exponential moving average (EMA) updates with decay coefficient $\tau = 0.996$ to provide stable masking guidance. This dual-ratio strategy encourages the model to rely more heavily on predicted missing representations during cross-modal consistency learning.

D.3 Implementation of Context-Aware Conditional Distribution Alignment

Computational considerations. The proposed context-aware 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 $\kappa(i, j)$ for all sample pairs scales as $\mathcal{O}(N^2)$ per batch. 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. To improve training stability, we treat the context representations $\mathbf{c}_i$ as constants when computing the kernel weights, i.e., gradients are not propagated through $\kappa(i, j)$. This prevents trivial solutions where the model adapts the context representations to

artificially minimize the distribution alignment loss. Additionally, we normalize the kernel weights for each query sample i to ensure numerical stability and bounded gradients in the conditional ECF estimation. Specifically, we use $L = 128$ random projection directions sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$, $K = 64$ frequency samples drawn uniformly from $[-t_{\max}, t_{\max}]$ with $t_{\max} = 5.0$, 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.

E USE OF LLMS

We employed large language models (LLMs) solely for polishing the writing. They were not used for other purposes, such as generating new ideas.