Learning Missing Modal Electronic Health Records with Unified Multi-modal Data Embedding and Modality-Aware Attention

Kwanhyung Lee[1*] KWANLEE9209@AITRICS.COM
Soojeong Lee[1,3†] DRLISA@AITRICS.COM
Sangchul Hahn[1] S.HAHN@AITRICS.COM
Heejung Hyun[1] ALEX.HYUN@AITRICS.COM
Edward Choi[2] EDWARDCHOI@KAIST.AC.KR
Byungeun Ahn[1] BEN@AITRICS.COM
Joohyung Lee[1*‡] CHRIS@AITRICS.COM
[1] *AITRICS Inc.*
[2] *Korea Advanced Institute of Science and Technology (KAIST)*
[3] *Sungkyunkwan University (SKKU)*

Abstract

Electronic Health Record (EHR) provides abundant information through various modalities. However, learning multi-modal EHR is currently facing two major challenges, namely, 1) irregular and asynchronous sampling and 2) modality missing. Moreover, a lack of shared embedding function across modalities can discard the temporal relationship between different EHR modalities. On the other hand, most EHR studies are limited to relying only on EHR Times-series, and therefore, missing modality in EHR has not been well-explored. Therefore, in this study, we introduce a Unified Multi-modal Set Embedding (UMSE) and Modality-Aware Attention (MAA) with Skip Bottleneck (SB). UMSE treats all EHR modalities without a separate imputation module or error-prone carry-forward, whereas MAA with SB learns missing modal EHR with effective modality-aware attention. Our model outperforms other baseline models in mortality, vasopressor need, and intubation need prediction with the MIMIC-IV dataset.

1. Introduction

Recently, electronic health record (EHR) emerges as a promising source of patient information. Utilizing its rich information, deep learning is making significant progress in various clinical regimes, especially in event prediction, e.g., mortality, sepsis, acute kidney injury, as well as the need for vasopressor administration, intubation, and ICU transfer (Sung et al. (2021); Wanyan et al. (2021)). Clinically, the early prediction of clinical events enables clinicians to effectively prioritize high-risk patients, allocate resources efficiently, and make prompt interventions (Choi et al. (2022)). Nevertheless, two problems in EHR are hindering

[*] Equal contribution
[†] Work done while S.Lee was an intern at AITRICS Inc.
[‡] Corresponding author

many promising deep learning algorithms to be readily transferred to learn EHR: missing modality and irregular/asynchronous sampling.

EHR encompasses a wide range of modalities, including not only EHR time series but also medical images (e.g., X-ray images), text (e.g., clinical notes, chief complaints), and demographics, all of which hold the potential for enhancing the predictive performance of clinical event (Lee et al. (2022); Hayat et al. (2022)). Among various modalities in EHR, time-series data is most frequently used for clinical event prediction, and many reported EHR studies solely rely on time-series EHR data (i.e., vital signs and laboratory test results) (Wanyan et al. (2021); Choi et al. (2022); Sung et al. (2021); Kim et al. (2019); Che et al. (2018); Shukla and Marlin (2019); Tipirneni and Reddy (2022)). Though the predictive performance from EHR time series can be improved by supplementing other modalities Lee et al. (2022), multi-modal learning has not been widely explored in EHR learning.

One of the major challenges in multi-modal EHR learning is the missing modality. Specifically, in practice, not all data modalities are consistently available for patients. Frequent modality missing in EHR data impedes the use of multi-modal fusion models to fuse a wide range of EHR data. Moreover, a weak inter-modality relationship and varying dimensionalities of different EHR modalities even complicate learning multi-modal EHR with missing modalities.

Various studies have addressed the missing modality problem. For example, many studies have approached the missing modality problem with generative methods (Ma et al. (2021); Vasco et al. (2020)), but the generative method is unsuitable for learning multi-modal EHR due to the excessive heterogeneity of modalities of EHR (e.g., time-series pulse data cannot generate X-ray images). Ma et al. (2022) introduced self-attention masking for missing text modality but reported performance degradation when trained with missing-modal data. Moreover, because they modeled the relationship between every possible token pair regardless of the token modality, the computational cost increases much with the increasing number of modalities ($O(n^2)$), which can thus be less scalable for multi-modal fusion. Hayat et al. (2022) utilized basic LSTM structure to late-fuse the X-ray imaging to the time-series EHR in mortality and phenotype prediction. LSTM can handle the missing modality problem, for LSTM can function with variable input length.

Other than the missing modality, a lack of shared embedding functions across EHR modalities can be problematic since a unified embedding method can model the temporal relationship between different modalities. Reported EHR embedding studies usually address irregularity/asynchrony in EHR time-series data only Horn et al. (2020); Choi et al. (2022); Tipirneni and Reddy (2022). However, other EHR modalities, e.g., medical images, clinical notes, and laboratory test results, are recorded at irregular time intervals as well, depending on factors such as clinical protocols, patient conditions, and healthcare settings (Che et al. (2018); Shukla and Marlin (2019); Tipirneni and Reddy (2022); Zhang et al. (2020)). In fact, Zhang et al. (2020) have shown that learning clinical notes without considering their occurrence time information can lead to misclassification. Yet, no single unified embedding function method for different modalities has been introduced. Therefore, in this study, we propose a unified set embedding that addresses the irregularity/asynchrony of all modalities without a separate imputation module. In summary, our contributions are:

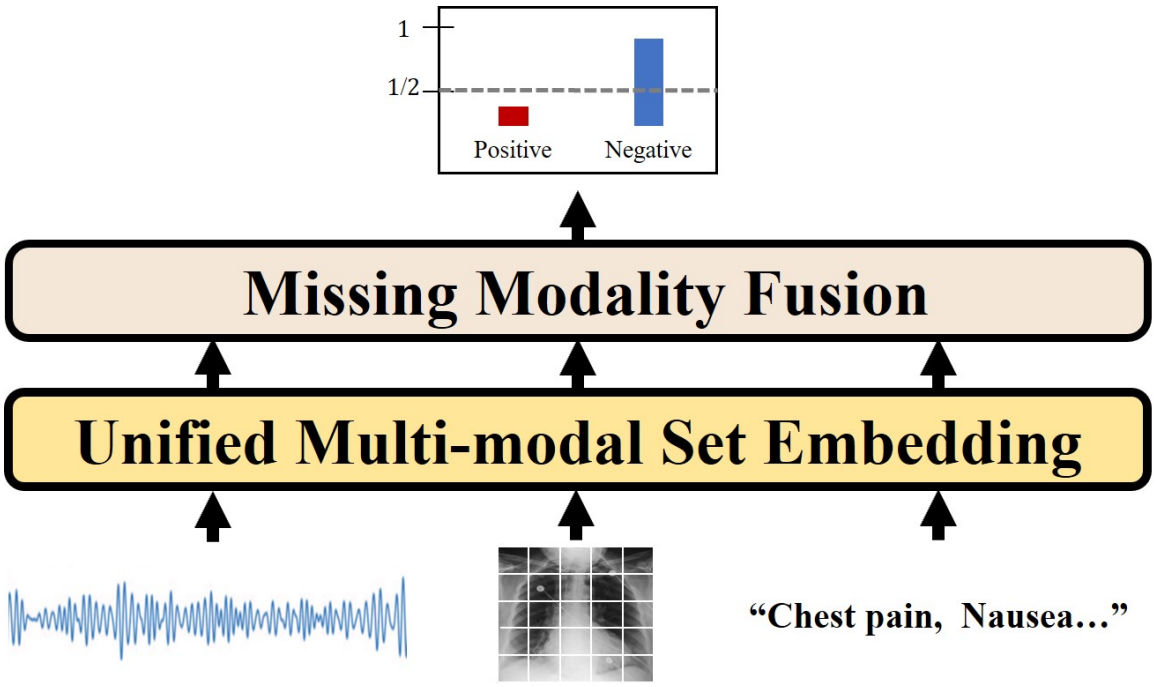

Figure 1: Overview of our proposed method consisting of 1) **Unified Multi-modal Set Embedding (UMSE)** which embeds accurate time and feature type information to all modality data and 2) **Missing Modality Fusion**.

- We propose a Unified Multi-modal Set Embedding (UMSE) as an efficient embedding method for Multi-modal EHR. UMSE views all modalities in the same line and provides a unified method to 1) solve irregular/asynchronous problems of all EHR modalities, 2) utilize the time information of all EHR modalities, and 3) model the temporal relationship between different modalities by sharing the time embedding function across different modalities.
- We suggest Modality-Aware Attention (MAA) and Skip-Bottleneck fusion (SB) to effectively learn multi-modal EHR with modality missing. MAA assigns distinct attention to each modality compared to the averaging method from Multi-modal Bottleneck Transformer (MBT) Nagrani et al. (2021) whereas SB enables MBT to learn with missing modality.
- We provide extensive analysis of each component of our proposed model to show the effects of different design choices. Here, we discuss 1) the application of UMSE on different modalities, 2) MAA comparison, 3) the effects of pretraining (Appendix 8) and 4) modality combinations.
- In this paper, we extensively experimented with three different clinical tasks: mortality, vasopressor need, and intubation need predictions with publicly open-access real-world large dataset Medical Information Mart for Intensive Care (MIMIC-IV),

MIMIC Chest X-ray (MIMIC-CXR) and MIMIC Emergency Department (MIMIC-ED) (Johnson et al. (2020, 2019, a)). With a real-time time embedding (Appendix B.1) for an online monitoring scenario using maximum 1440 hours of each subject, we estimated the performance in a practical setting.

- Lastly, considering the lack of benchmark in the field of EHR multi-modal fusion with missing modality while handling irregular sampling, we release our code to ensure the reproducibility and applicability of our approach: https://github.com/AITRICS/Medical_Tri_Modal_Pilot.git.

Generalizable Insights about Machine Learning in the Context of Healthcare

Learning multi-modal EHR usually involves two problems: 1) irregular/asynchronous data and 2) missing modality. Though many studies have proposed to solve irregular/asynchronous data problems, they are limited to apply their method for time-series EHR. However, Zhang et al. (2020) have demonstrated the need for the time information on EHR Text as well, and a lack of shared embedding method across modalities can discard the important temporal relationship between different EHR modalities. On the other hand, Lee et al. (2022) have shown that multi-modal bottleneck fusion outperforms other regressors such as vanilla Transformer in prediction tasks using EHR. However, bottleneck fusion (BF) has two drawbacks to be readily applied to learn multi-modal EHR: 1) BF does not allow missing modality, 2) BF neglects different importance of modalities since it computes the final logit by averaging logits from each modality. To tackle the aforementioned problem, we suggest a Unified Multi-modal Set Embedding (UMSE), Skip Bottleneck (SB), and Modality-Aware Attention (MAA). First, UMSE effectively handles irregular/asynchronous data problems of all modalities, and it can encourage modeling the temporal relationship between different modalities by sharing the time embedding function across all modalities. Second, SB empowers bottleneck fusion to handle missing modalities. Lastly, MAA employs modality-aware attention scores to compute the final logit.

2. Related Works

2.1. Learning Multi-Modal EHR for Event Prediction

Efforts to obtain a more comprehensive understanding of patient patterns for accurate clinical event prediction have led numerous researchers to employ multi-modal EHR datasets. Many have reported clinical event prediction methods using various EHR data combinations such as time-series EHR combined with clinical text (Lee et al. (2022); Wang and Lan (2022); Suresh et al. (2017); Qin et al. (2021); Lyu et al. (2022)), medical codes (e.g., procedure code, diagnosis code) alongside EHR text (Qiao et al. (2019)). Choi et al. (2022) have incorporated time-series vital signs with wearable device-based heart signal data for clinical event prediction. Meanwhile, Vale-Silva and Rohr (2020) have employed patient genetics, clinical, and histopathology slide images for long-term pan-cancer survival prediction.

2.2. Multi-Modal Fusion

Numerous research methodologies have been investigated to obtain extensive knowledge from multi-modal data. To learn both tri-modal and bi-modal interactions, Zadeh et al. (2017) employed the Tensor Fusion method. Several studies have utilized the Transformer architecture (Vaswani et al. (2017)) for multi-modal fusion; Kim et al. (2021) adopted Transformer structure to integrate multi-modal data, without requiring encoders dedicated to each modality; Tsai et al. (2019) used cross-modal attention module to learn varied bi-modal interactions of the multi-modal data; Akbari et al. (2021); Alayrac et al. (2020) applied self-supervision techniques to train modality-specific encoders to project their modality representations to a common space dimension for improved downstream performance in learning video/audio/text; and Nagrani et al. (2021) employed bottleneck learnable tokens to model inter-modality interaction with the Transformer architecture without much-increasing computation burden.

2.3. Missing Modality

In practice, it is often the case that not all modalities are available for every patient, leading to the issue of "missing modality". Intuitively, many researchers have explored methods for handling missing modalities by generating representative vectors. Ma et al. (2021); Vasco et al. (2020) learned to generate representative vectors of missing modalities. Poklukar et al. (2022) trained encoders to make representations of all missing modality combinations similar to the representation of the full modality combination. Ma et al. (2022) simply employed a masking method for self-attention to address missing modality and suggested layer fusion with multi-task learning (multi-token) to enhance model robustness to missing modality. Hayat et al. (2022) used LSTM to fuse modality-wise representations with possible modality-missing. Vale-Silva and Rohr (2020) filled missing feature values with median substitution.

2.4. Data Embedding in EHR

Various embedding strategies have been reported to incorporate the occurrence time of EHR data. Che et al. (2018) utilize each feature's missing period as their temporal information. Choi et al. (2022); Lee et al. (2022); Hayat et al. (2022) used 1 or 2 hourly sampling method to discretize the temporal axis of time series EHR data. Horn et al. (2020) suggested a set encoding method to train the model with irregularly sampled time-series EHR data without carry-forward or separate generative module. Tipirneni and Reddy (2022) employed learnable set embeddings in predicting mortality from EHR time-series data. Zhang et al. (2020) devised Flexible Time-aware Long Short Term Memory (FT-LSTM) in order to use both time information and hierarchical information of clinical text.

3. Methods

In this section, we describe our method to effectively learn multi-modal EHR data with modality missing. Our model aims to tackle two problems in learning multi-modal EHR, namely, 1) EHR data embedding, and 2) modality fusion with missing modality. As depicted in Figure 1, our Unified Multi-modal Set Embedding (UMSE) and Missing Modality

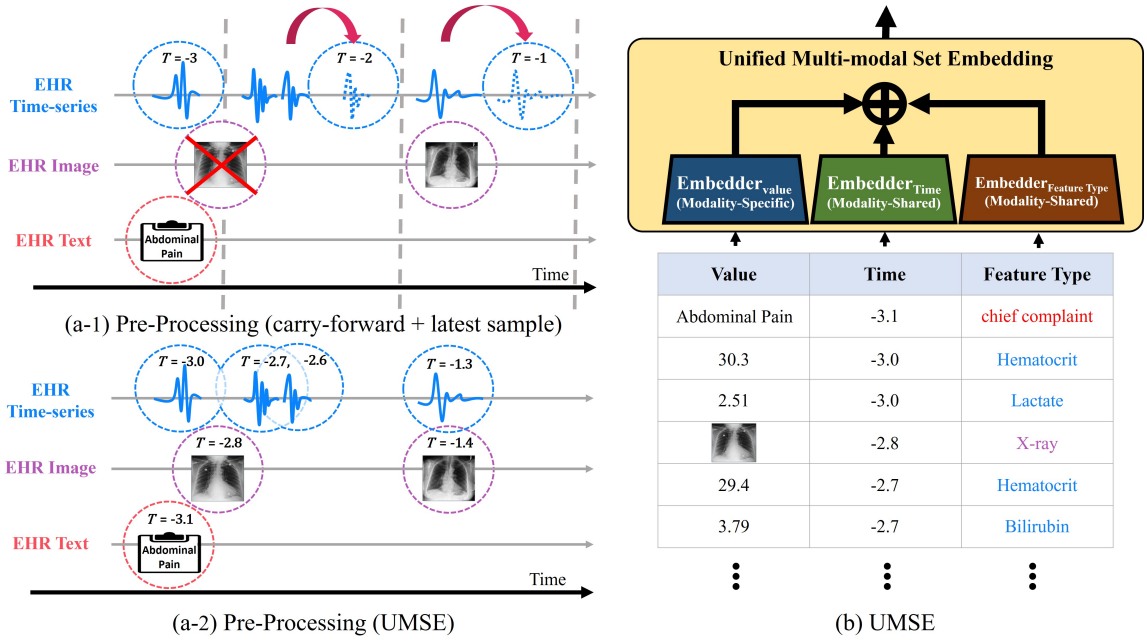

Figure 2: (a-1) Overview of traditional preprocessing strategy for EHR multi-modal data: carry-forward for regular time-grid and latest sampling for non-Time-series modality (Lee et al. (2022); Hayat et al. (2022)). (a-2, b) Our Unified Multi-modal Set Embedding (UMSE).

Fusion module (MMF) address these problems. Specifically, UMSE allows using irregular/asynchronous multi-modal EHR data with neither a separate imputing module nor loss of time information. Moreover, UMSE can model inter-modality temporal relationships by sharing the time embedding function across modalities. On the other hand, MMF enables the model to learn multi-modal data with possible modality-missing through Skip Bottleneck (SB). Moreover, through Modality-Aware Attention (MAA), MMF assigns different attention scores for each modality to enhance the predictive performance. All three prediction tasks (i.e. mortality, vasopressor need, and intubation need prediction) are binary classification tasks predicting whether the event would occur within 12 hours.

3.1. Unified Multi-modal Set Embedding

Our Unified Multi-modal Set Embedding (UMSE) aims to tackle 1) the irregular and asynchronous nature of EHR multi-modal data, and 2) inter-modality temporal relationships. In practice, both irregularity and asynchrony occur in time-series EHR, whereas only irregularity exists in other EHR modalities such as EHR Image. Traditionally, error-prone carry-forward or separate imputation modules have been widely employed for the time-series EHR, and for other EHR modalities, time information is usually discarded.

Definition 1 (Irregularity). *We consider an arbitrary EHR feature B occurs N times, i.e. $B := \{(S_1, t_1), ..., (S_N, t_N)\}$, where t_n denotes the occurrence time of the feature D with its value S_n and $t_i < t_{i+1}$. A feature B is irregularly sampled if there exists at least one t_i such that $t_{i+1} - t_i \neq t_i - t_{i-1}$.*

Definition 2 (Asynchrony). *A D-dimensional EHR feature B occurs N times, i.e., $B := \{b_1, , ..., b_N\}$. A feature B is asynchronous if there exists at least one b_i at which at least one element is missing, i.e., $|b_i| \neq D$.*

Definition 3 (Multi-modal EHR). *We denote a multi-modal EHR data of i^{th} subject as a set S_i of $N := |S_i|$ observations s_i where $S_i := \{s_1, ..., s_N\}$. We treat each observation s_i as a triplet (v_i, t_i, FT_i), consisting of an observed value $v_i \in \mathcal{R}^{M_{FT_i}}$, observation time $t_i \in \mathcal{R}$, observed feature type indicator $FT_i \in \{1, ..., D\}$, where D represents the dimensionality of the whole multi-modal EHR including not only EHR image, EHR text, but also each feature in EHR Time-series, e.g., Hematocrit, Lactate, etc. $M_{FT_i} = 1$, 224×224 when FT_i is EHR Time-series, EHR image, respectively. For EHR text, v_i is not numeric but text string.*

Inspired by Horn et al. (2020); Tipirneni and Reddy (2022), our UMSE rephrases the problem of encoding multi-modal EHR into the problem of encoding a set of observations as described in the above definition. To this end, our UMSE consists of three embedding functions as illustrated in Figure 2: value embedder, time embedder, and feature type embedder, which are denoted as $Embedder_{Value}$, $Embedder_{Time}$, $Embedder_{FeatureType}$, respectively. These three embedders encode each element of the observed triplet and add them up. The output of the UMSE is then concatenated with the outputs from other observations.

We use a modality-specific encoder for the value embedder. Specifically, we use pre-trained frozen Swin-Transformer Liu et al. (2021) followed by a linear projection for the EHR image. For EHR text, we used a BERT tokenizer and pre-trained & frozen BioBERT Lee et al. (2020) followed by a linear projection. For EHR Time-series, we used a simple linear projection with nonlinearity. Please refer to Appendix A.5.2, A.4.3 for more details.

The dimension of the input/output of the value embedder varies with the modality, i.e., EHR Time-series, EHR image, and EHR text. Specifically, $Embedder_{Value} : \mathcal{R} \rightarrow \mathcal{R}^{256}$, for EHR Time-series, $Embedder_{Value} : \mathcal{R}^{224 \times 224} \rightarrow \mathcal{R}^{49 \times 256}$ for EHR image, $Embedder_{Value} : \mathcal{S} \rightarrow \mathcal{R}^{128 \times 256}$, for EHR text ($\mathcal{S}$ denotes string).

Time embedder and feature type embedder are shared across all modalities to model inter-modality temporal relationships. Inspired by Tipirneni and Reddy (2022), we use a look-up table and simple linear projection with nonlinearity for the feature type embedder and time embedder, respectively.

3.2. Missing Modality Fusion

Our Missing Modality Fusion module (MMF) is depicted in Figure 1 and in Figure 3 models the intra-modality and inter-modality interaction for three different clinical prediction tasks, i.e., mortality, vasopressor need, and intubation need prediction. Our MMF consists of two modules, i.e., Skip Bottleneck (SB) and Modality-Aware Attention (MAA). SB enables the Multimodal Bottleneck Transformer (MBT) to handle data with missing modality whereas MAA provides modality-wise attention to consider the logit from differ-

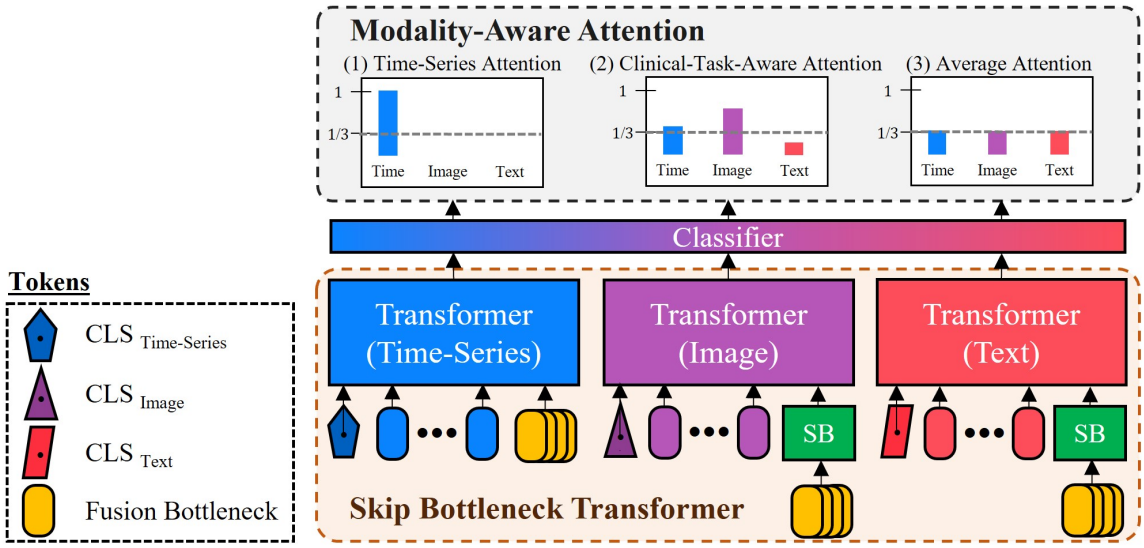

Figure 3: Overview of our Missing Modality Fusion module (MMF) consisting of Skip Bottleneck (SB) Transformer with three different modality-aware attention schemes: (a) Time-Series Attention (TSA), which focuses on Time-series modality only, (b) Clinical-Task-Aware Attention, which dynamically attends modalities depending on target task and observing modalities, (c) Average Attention (AA), which considers all modalities equally. Note that TSA, CTAA, and AA are built on top of MBT with SB. Note that AA is simply MBT with SB.

ent modality-transformer differently. All of our MAA processes are applied on top of the MBT with SB for cases involving missing modalities.

3.2.1. Revisiting Bottleneck Fusion

Nagrani et al. (2021) proposes the MBT (Multimodal Bottleneck Transformer) architecture, which effectively reduces the computational costs of transformer models. MBT enables modality interaction exclusively via fusion bottleneck tokens shared among modality-specific transformer layers (Equation 1). This structure encourages intra-modality interaction while managing inter-modality interaction through a narrow bottleneck, which may be advantageous for EHR multimodal data due to its inherent heterogeneity and rather weak correlations. In a multi-modal transformer, Ma et al. (2022) simply used the attention masking method to handle missing modality during inference. The masked attention excludes missing modality in self-attention softmax calculation which prevents unnecessary interaction between observed modality data and unobserved missing modality data.

3.3. Skip Bottleneck Transformer

Lee et al. (2022) have shown that Multimodal Bottleneck Transformer (MBT) Nagrani et al. (2021) outperforms Multimodal Transformer (MT) Ma et al. (2022) and other regressors in multi-modal EHR learning. However, in practice, a single mini-batch comprises subjects

with various combinations of modalities, and MBT cannot process a mini-batch with varying modalities. To supplement MBT for missing modality, we propose Skip Bottleneck (SB).

SB consists of two simple processes; 1) we feed random numbers (e.g. zero-vectors) to the Transformer of the missing modalities; 2) we 'skip' the bottleneck tokens from the missing modality for fusion (Equation 2) (Figure 3).

$$|z_i^{l+1}||\hat{z}_{fsn_i}^{l+1}| = Transformer(|z_i^l||z_{fsn}^l|;\theta_i) \tag{1}$$

$$\begin{aligned} z_{fsn}^{l+1} &= \frac{1}{1+\mathbb{1}_{Image}+\mathbb{1}_{Text}}(\hat{z}_{fsn_{Time-Series}}^{l+1} + \mathbb{1}_{Image}\hat{z}_{fsn_{Image}}^{l+1} + \mathbb{1}_{Text}\hat{z}_{fsn_{Text}}^{l+1}), \\ \mathbb{1}_i &= \begin{cases} 1, & \text{if modality i is present} \\ 0, & \text{if modality i is NOT present} \end{cases} \end{aligned} \tag{2}$$

z_i^l refers to the token at the Transformer layer l of the observed modality i, whereas $\hat{z}_{fsn_i}$ and z_{fsn} refers to the bottleneck fusion token before and after averaging, respectively. As illustrated in Figure 3 and Equation 2, the temporary bottleneck token $\hat{z}_{fsn_i}$ from EHR Time-series Transformer is never skipped since subjects in multi-modal EHR always possess Time-series data.

3.4. Modality-Aware Attention Decision Making

The original MBT (Nagrani et al. (2021)) averages the pre-softmax logits of CLS_M tokens before feeding them to the shared classification layer. This approach inherits the assumption that all modalities are equally significant in the decision-making process for visual-audio tasks. However, in EHR multi-modal learning, time-series data is often regarded to be more significant than other modalities. Consequently, we added two more designs to the traditional Average Attention (AA) to experiment with different modality attention: 1) Time-Series Attention (TSA), and 2) Clinical-Task-Aware Attention (CTAA) as described in (Figure 3). Specifically, TSA places the [CLS] token solely for the Time-series Transformer, while CTAA employs learnable scalars with temperature τ. As a result, CTAA creates a modality-wise attention score through the softmax function and determines which modality's logit should be prioritized (Equation 3). Note that all TSA, CTAA and AA are built on top MBT with SB with different MAA strategies.

$$Attention_m = \frac{exp(w_m/\tau)}{\sum_j exp(w_j/\tau)} \tag{3}$$

where w are three learnable logits of softmax and τ is the temperature to sharpen the softmax function to get a pre-sigmoid-logit before binary cross-entropy calculation. Note that m is a modality indicator.

4. Experiments

4.1. Dataset

In this study, we use three EHR datasets: MIMIC-IV[1], MIMIC-CXR[2], and MIMIC-ED[3] (Johnson et al. (2020, 2019, a)). As they share the same patients, we merged them to obtain per-patient information of vital signs, lab results, demographics from MIMIC-IV, X-ray images from MIMIC-CXR, and chief-complaint text from MIMIC-ED. Additionally, we used the MIMIC-IV-Note[4] dataset for clinical notes (Johnson et al. (b)) and compared the performance when it replaces the chief-complaint in Appendix B.4 with a data table Appendix A.4. Unless otherwise specified, EHR text data refers to chief-complaint from MIMIC-ED throughout this study. If chief-complaint text data exists, it means that the patient has visited ED before ICU admission. Details are provided below.

Table 1: Data statistics with the number of subjects for mortality prediction, vasopressor need, and intubation need prediction tasks with modality missing information.

(a) Mortality Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	3486	30870	430	3741	413	3873
Image Missing Rate	75.22%	77.84%	73.02%	78.56%	74.09%	78.23%
Text Missing Rate	46.76%	50.61%	46.28%	51.70%	43.83%	50.40%
(b) Vasopressor Need Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	9341	24822	1172	2969	1183	3085
Image Missing Rate	73.30%	79.16%	73.29%	79.69%	74.64%	79.00%
Text Missing Rate	49.33%	50.61%	48.63%	52.24%	47.08%	50.86%
(c) Intubation Need Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	13450	20682	1665	2467	1716	2552
Image Missing Rate	72.71%	80.71%	73.51%	80.91%	73.66%	80.60%
Text Missing Rate	56.71%	46.09%	58.62%	46.25%	55.48%	45.96%

4.1.1. Data Preprocessing

- **EHR Times-series and Demographic data**: For each ICU patient, we collected EHR numeric data from MIMIC-IV, ranging from a minimum of 3 hours to a maximum of 1440 hours (60 days). Numeric data comprises demographic features (age and gender), as well as time-series data, i.e., vital signs and lab-test results. Vital sign includes six features: heart rate, respiration rate, diastolic and systolic blood pressure, temperature, and pulse oximetry. The laboratory result data, i.e. lab-test, encompasses ten features: Hematocrit, Platelet, Bilirubin, etc, following Sung et al.

1. https://physionet.org/content/mimiciv/1.0/
2. https://physionet.org/content/mimic-cxr/2.0.0/
3. https://physionet.org/content/mimic-iv-ed/2.2/
4. https://physionet.org/content/mimic-iv-note/2.2/

(2021) (See Appendix A.3). In total, there are 18 numeric data features. We exclude patients without (or less than) 5 vital-sign features during the entire ICU hospitalization period. We applied min-max normalization using our training set. More detailed information regarding our EHR time-series data is provided in Appendix A.3.

- **EHR Text data**: We extract chief-complaint text from MIMIC-ED and admission-related text (Chief Complaint, Medication on Admission, Past Medical History) from MIMIC-IV-Note. We described more detailed information about EHR text data preprocessing steps in Appendix A.4.
- **X-ray Image data**: We preprocess MIMIC-CXR X-ray image by removing the black margin area and excluding images with aspect ratios bigger or less than 1.3 or 0.7 respectively. More detailed image preprocessing information, pre-training strategy, and image augmentation method are described in Appendix A.5.

For training, we randomly selected a time window ranging from 3 to 24 hours to predict the occurrence of clinical events within the next 12 hours. We excluded any time windows with no EHR Time-series data within the most recent hour interval, which is the latest 1-hour within the training time windows, for both training and inference. During training, we extracted positive and negative windows with equal ratios using a batch sampler as described in Appendix A.2. For inference, we randomly selected and fixed 5 positive and 5 negative periods per patient during ICU hospitalization period.

4.1.2. Data Split

We randomly selected 80%, 10%, and 10% of patients for training, validation, and test set. For each patient, we extracted EHR Time-series data with EHR Text data and EHR image data from MIMIC-ED and MIMIC-CXR with date time information indicating when the text or image was captured. Table 1 illustrates that not all modalities are paired for each sample, and provides missing rate information.

4.2. Clinical Objectives

We extracted three tasks-related information, i.e., mortality, vasopressor need, and intubation need predictions with the following statistics.

- **Mortality prediction within 12 hours**: As depicted in Table 1-(a), we utilized 42,813 ICU cases, comprising 4,329 positive cases with defined mortality onset times.
- **Vasopressor need prediction within 12 hours**: As depicted in Table 1-(b), we utilized 42,572 ICU cases, including 11,696 positive cases labeled with vasopressor initiation times. Labels were assigned when Norepinephrine, Dopamine, Dobutamine, or Epinephrine was administered. The Appendix A.1.1 contains item number details.
- **Intubation need prediction within 12 hours**: As outlined in Table 1-(c), we extracted 42,532 ICU cases from which 16,831 cases were labeled positive with intubation start times. We focused on 7 intubation types among diverse MIMIC-IV chart events. Specific intubation types and item numbers are detailed in Appendix A.1.1.

4.3. Baseline Models

We compare the performance of our model using test set AUPRC and AUROC with previously reported EHR multi-modal algorithms. Since our prediction tasks are highly imbalanced, we primarily focus on AUPRC and considered AUROC secondarily. To ensure a fair comparison, all classification layers in this paper consist of a 2-layer Multi-Layer Perceptron (MLP) with Layer Normalization (LN) and ReLU non-linearity between the two linear layers. We process EHR static features, i.e., age and gender, through one linear projection with ReLU nonlinearity; we concatenate it to the output of all fusion algorithms (e.g., Transformer). All transformer fusion networks have 6 layers with 256 feature dimensions. We conducted a learning rate sweep ranging from 10^{-6} to 10^{-4}. All models are trained with AdamW optimizer, 50 epochs with a 3-seed averaging. We compare our model performance against the following algorithms:

- **HAIM**: HAIM (Soenksen et al. (2022)) used varying pre-trained modality-specific encoders for each modality. The encoder outputs are then concatenated before linear classification layers. For missing modalities, a zero-padding strategy is employed.
- **MNRIFN**: Proposed by Wang and Lan (2022), this method only accommodates sequential multi-modal data; we reproduced using our time-series and text data.
- **Medfuse**: Hayat et al. (2022) used a LSTM to fuse bi-modal EHR data with modality missing case samples. In this paper, We reproduced this model as: 1) Bi-modal with Time-series and Images, and 2) Tri-modal with Time-series, Text, and Images.
- **Multi-modal Transformer (MT)**: Ma et al. (2022) utilized the Transformer architecture with attention mask to fuse image and text with missing modality.
- **Multimodal Bottleneck Transformer(MBT)**: Nagrani et al. (2021) devised a modality-wise Transformer with bottleneck token to model inter-modality interaction. Since the original MBT can not receive data with modality missing, we develop our Skip Bottleneck (SB) to MBT for data with missing modality.

We used carry-forward imputation for HAIM, MNRIFN, Medfuse, and MT multi-modal input data. Given UMSE embedding's compatibility with only transformer architectures, we applied UMSE to MT and MBT only.

4.4. Robustness Against Missing Modality

We evaluate the model robustness against missing modality. To do so, we assess the model performance not only with the original test set but also with the test set with an increasing modality missing rate. In this study, we conduct three different experiments: 1) missing robustness of MBT with three different modality-attention scores, 2) training strategies to increase model robustness against modality missing, and 3) UMSE on missing robustness. First, we assess the missing robustness of TSA, CTAA, and AA as illustrated in Figure 4. Second, we explore two different strategies to increase missing robustness, i.e., Missing-Modal Augmentation (MMA) and layer optimization (Table 4) with multi-task learning (i.e., multi-token) as suggested in Ma et al. (2022). For layer optimization, we varied the fusion

starting layer of TSA and selected the best fusion starting layer, which we call Fusion Layer Search (FLS), based on validation AUPRC. Moreover, we implemented the multi-token strategy as suggested by Ma et al. (2022). Lastly, we examined if UMSE contributes to the missing robustness based on TSA.

5. Results

In this section, we report the predictive performance of 1) different multi-modal fusion models (Section 5.1) and 2) various strategies to enhance the robustness against missing modality (Section 5.3, 5.4) using the test AUPRC and AUROC, averaged over 3 runs.

Table 2: Performance comparison between baseline models and our proposed models for three clinical event prediction tasks, using bimodal (a, b) and trimodal (c) data.

(a) EHR Times-Series + EHR Text

	Mortality		Vasopressor Need		Intubation Need	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
MNRIFN	0.703±0.006	0.930±0.001	0.649±0.005	0.863±0.001	0.507±0.006	0.749±0.003
HAIM	0.704±0.003	0.933±0.002	0.691±0.0	0.880±0.001	0.572±0.001	0.800±0.001
MT+carry-forwrd	0.686±0.007	0.921±0.002	0.667±0.007	0.869±0.001	0.579±0.001	0.800±0.0
MT+UMSE	0.787±0.005	0.956±0.001	**0.751±0.001**	**0.906±0.001**	0.739±0.012	0.902±0.005
ours (AA)	**0.801±0.004**	**0.961±0.002**	0.743±0.002	0.901±0.002	0.744±0.009	0.903±0.005
ours (TSA)	0.790±0.006	0.955±0.002	0.748±0.002	0.903±0.002	**0.751±0.002**	**0.904±0.001**
ours (CTAA)	0.789±0.004	0.955±0.002	0.748±0.002	0.904±0.001	0.751±0.001	0.903±0.002

(b) EHR Times-Series + EHR Image

	Mortality		Vasopressor Need		Intubation Need	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Medfuse	0.626±0.005	0.924±0.003	0.698±0.004	0.884±0.001	0.583±0.002	0.802±0.001
HAIM	0.696±0.007	0.927±0.004	0.703±0.001	0.885±0.0	0.579±0.002	0.802±0.001
MT+carry-forward	0.715±0.004	0.927±0.006	0.732±0.002	0.893±0.001	0.682±0.006	0.831±0.004
MT+UMSE	0.799±0.005	0.960±0.002	0.778±0.005	0.915±0.004	0.794±0.006	0.908±0.0
ours (AA)	0.807±0.006	0.963±0.003	0.78±0.007	0.915±0.0	0.803±0.004	0.913±0.002
ours (TSA)	**0.811±0.01**	**0.963±0.003**	0.777±0.003	0.916±0.002	**0.807±0.001**	**0.915±0.001**
ours (CTAA)	0.805±0.001	0.962±0.002	**0.781±0.002**	**0.916±0.001**	0.804±0.006	0.915±0.0

(c) EHR Times-Series + EHR Text + EHR Image

	Mortality		Vasopressor Need		Intubation Need	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Medfuse	0.819±0.007	0.955±0.002	0.800±0.001	0.912±0.001	0.775±0.001	0.871±0.0
HAIM	0.699±0.007	0.931±0.003	0.687±0.004	0.881±0.002	0.575±0.005	0.798±0.001
MT+carry-forward	0.784±0.001	0.929±0.004	0.781±0.005	0.915±0.004	0.764±0.0	0.862±0.001
MT+UMSE	0.844±0.004	0.961±0.005	0.831±0.003	0.926±0.001	0.854±0.003	0.927±0.001
ours (AA)	0.847±0.002	0.965±0.003	0.831±0.006	0.926±0.004	0.861±0.003	0.931±0.0
ours (TSA)	**0.864±0.01**	**0.970±0.005**	0.836±0.001	0.929±0.001	0.860±0.003	0.931±0.0
ours (CTAA)	0.854±0.005	0.964±0.001	**0.837±0.004**	**0.928±0.004**	**0.868±0.003**	**0.934±0.002**

5.1. Comparison with State-of-the-art

Table 2 shows that our proposed models (i.e., TSA, CTAA, as described in Figure 3) score the highest predictive performance in all three clinical tasks. Specifically, TSA and CTAA exhibit the highest performance (AUPRC/AUROC) in tri-modal fusion; 0.864/0.970, 0.837/0.928, 0.868/0.934 in mortality, vasopressor need, and intubation need prediction. Moreover, TSA outperforms CTAA in mortality prediction, which may show the importance

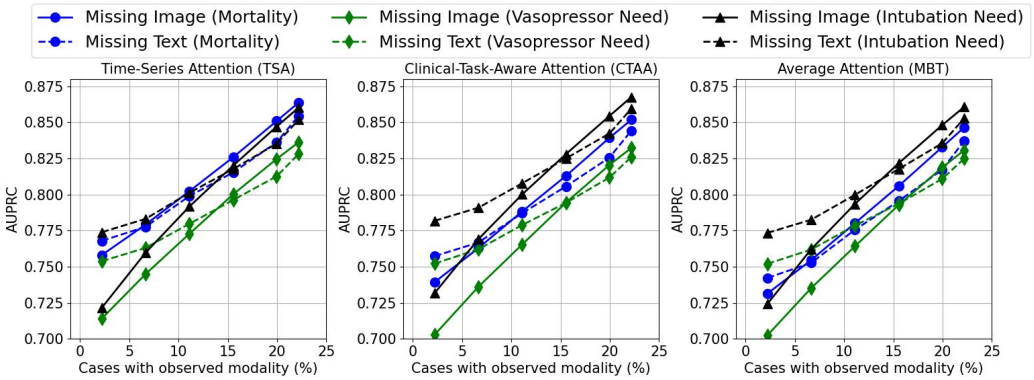

Figure 4: Missing modal robustness with different modality-aware attention methods.

of EHR Time-series in mortality prediction. Note that EHR Image usually scores higher predictive performance than EHR Text with EHR Time-series, indicating that EHR Image provides more supplementary information for EHR Time-series in all predictive tasks. In addition, UMSE clearly improves the performance, whereas the performance benefit of our fusion model over MT is relatively incremental.

Table 3: Comparative AUPRC performance analysis of MT and TSA when applying carry-forward or UMSE

	Mortality	**Vasopressor Need**	**Intubation Need**
MT+carry-forward	0.78±0.0	0.78±0.01	0.76±0.0
MT+UMSE	0.84±0.0	0.83±0.0	0.85±0.0
TSA+carry-forward	0.79±0.01	0.78±0.0	0.77±0.01
TSA+UMSE	**0.86±0.01**	**0.84±0.0**	**0.86±0.0**

5.2. Set embedding benefits not only vanilla transformer but also Time Series Attention

Table 3 demonstrates the performance superiority of UMSE over the most conventional alternative, the carry-forward, when using the vanilla transformer, MT, to fuse multimodal EHR with missing modality. Moreover, the performance improvement when UMSE is applied to MT and TSA is similar.

5.3. Missing EHR Image decreases the predictive performance more than missing EHR Text

We compare the robustness against missing modality for three different modality-aware attention strategies (Figure 4). Note that the case percentage (x-axis) is the ratio from the whole test cases (i.e., 4286, 4268, 4268, see Table 1). As described in Table 1, the number of test cases with EHR Image is approximately 25% in all three clinical tasks, and therefore, the maximum case percentage is approximately 25%. When an equal number of cases lose their EHR Image and EHR Text, we can observe that losing EHR Image

decreases the predictive performance more. Moreover, among three different clinical tasks, intubation need prediction may be more vulnerable to losing EHR Image indicated by the largest decaying slope for increasing EHR Image missing rate.

5.4. The optimal strategy for missing modality varies with the clinical tasks

We compare two different strategies to increase the robustness against the modality missing (Figure 5): 1) Missing-Modality Augmentation (MMA) and 2) fusion layer optimization with multi-task learning as suggested in Ma et al. (2022). As shown in Figure 5, no single strategy excels. MMA outperforms others in missing EHR image for mortality and intubation need prediction, whereas Ma et al. (2022) outperforms others in vasopressor need for missing either modality. It has to be noted that missing EHR text for mortality and intubation need prediction was not improved by either strategy.

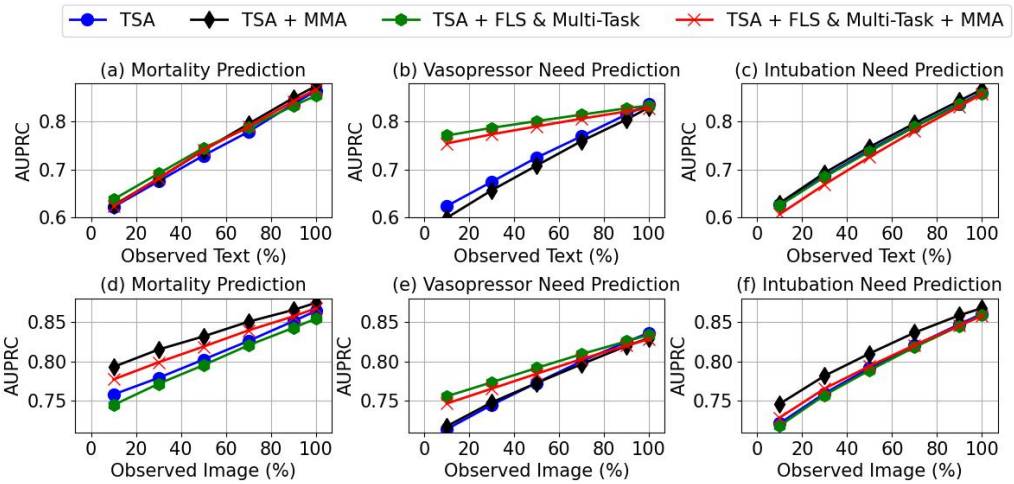

Figure 5: Result of the robustness against missing modalities for various training strategies, i.e., MMA, FLS with Multi-Task, and Both.

6. Discussion

6.1. Both Skip Bottleneck and Modality-Aware Attention modules improve learning multi-modal EHR

All of our models (i.e., TSA, CTAA, AA) are built on top of Skip Bottleneck since MBT does not consider modality missing. According to Table 2, all our models outperform other algorithms (including MT) except in vasopressor need prediction with bi-modal EHR. Our additional modality-aware attention scheme (TSA, CTAA) further improves the predictive performance, especially in mortality prediction. Note that the mortality prediction depends more heavily on EHR Time-series than other modalities and clinical tasks as illustrated in Figure 7. We assume that the Skip Bottleneck with the default attention from MBT (AA) was outperformed by models with additional modality-aware attention scheme especially in

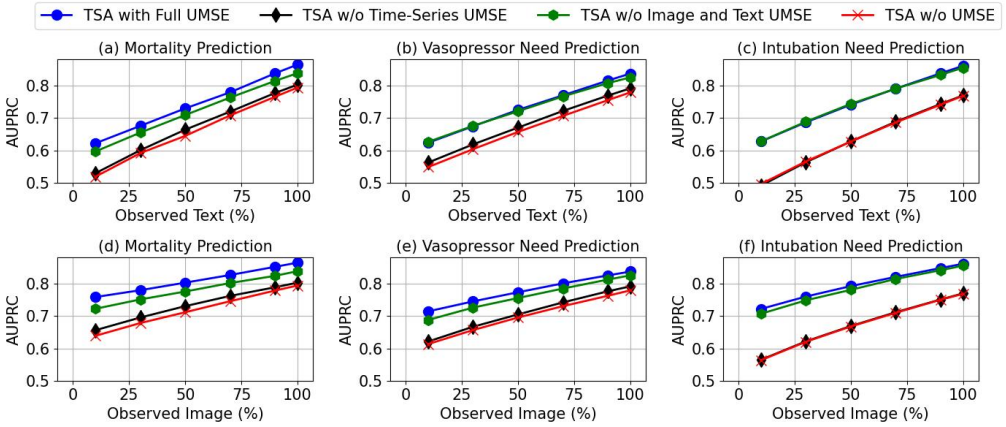

Figure 6: Result of the robustness against missing modalities for modality-wise UMSE implementation.

mortality prediction because mortality prediction is the clinical task with the most biased modality dependency as illustrated in Figure 7. Note that according to Figure 7, vasopressor need prediction task does not depend on a single modality whereas the intubation need prediction task makes slightly more attention to EHR Text than to other modalities.

Table 4: Result of optimal fusion layer search using TSA. The performance is measured in AUPRC and AUROC of validation dataset.

	Mortality		**Vasopressor Need**		**Intubation Need**	
L_{fusion}	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
1	0.861±0.007	0.968±0.003	0.843±0.003	0.929±0.001	0.862±0.001	0.932±0.001
2	0.858±0.006	0.967±0.002	0.842±0.003	0.928±0.003	0.865±0.003	0.933±0.0
3	**0.862±0.005**	**0.966±0.004**	0.841±0.001	0.928±0.002	**0.866±0.004**	**0.933±0.001**
4	0.858±0.003	0.968±0.002	0.840±0.002	0.927±0.003	0.864±0.005	0.932±0.003
5	0.857±0.007	0.969±0.0	**0.843±0.001**	**0.928±0.0**	0.865±0.001	0.933±0.002
6	0.818±0.007	0.965±0.002	0.766±0.002	0.913±0.001	0.759±0.002	0.907±0.002

6.2. Late fusion strategy can be detrimental to learning multi-modal EHR

As illustrated in Table 2, the Transformer-based fusion algorithms, which are all early fusion by default, excel other alternatives in all three clinical tasks (all other alternatives adopt a late fusion strategy). Moreover, in Table 4, the late fusion strategy drastically decreases the predictive performance in all clinical tasks. From these two observations, we can conjecture that late fusion is detrimental to fusing multi-modal EHR data.

6.3. Unified Multi-modal Set Embedding benefits multi-modal EHR learning

In addition to Table 3, Figure 6 also demonstrates that UMSE consistently enhances EHR multi-modal learning in all clinical tasks. This finding highlights that Unified Multi-modal Set Embedding (UMSE) method is advantageous not only for learning EHR time-series

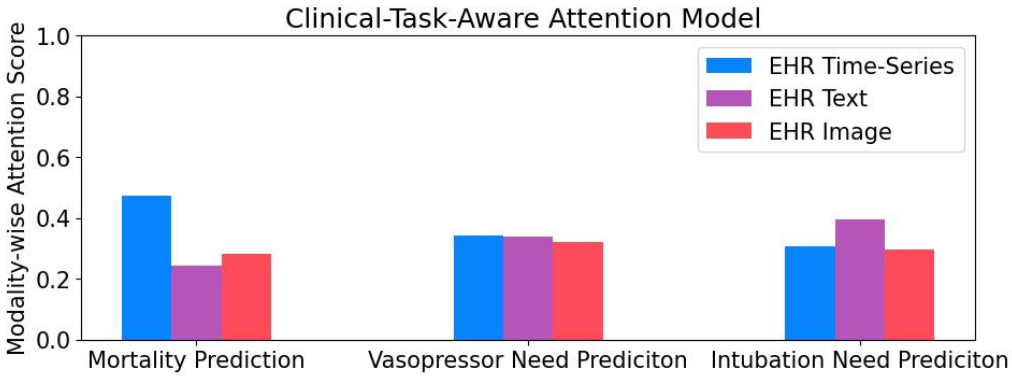

Figure 7: Modality-wise attention score from our proposed CTAA model.

data but also for modeling other EHR modalities such as EHR Images and EHR Text. The performance gain is consistent across all three clinical tasks and all missing modality rates.

6.4. Limitations

Our main focus is on 1) a unified set embedding for all EHR modalities and 2) modality-aware attention. We assess the robustness of our algorithms against missing modality (Figure 4, Figure 5, and Figure 6) with recently introduced approaches to tackle missing modality Ma et al. (2022). However, we hardly found a single solution to improve the modality-missing problem. We also discovered that no approach could improve the robustness on missing EHR Text in mortality and intubation need prediction. As a result, we believe that the research on a missing modality will be fitting to our following research.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovi'c, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *ArXiv*, abs/2006.16228, 2020.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Arom Choi, Kyungsoo Chung, Sung Phil Chung, Kwanhyung Lee, Heejung Hyun, and Ji Hoon Kim. Advantage of vital sign monitoring using a wireless wearable device for predicting septic shock in febrile patients in the emergency department: A machine learning-based analysis. *Sensors*, 22(18):7054, 2022.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. *arXiv preprint arXiv:2207.07027*, 2022.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. Mimic-iv-ed. a.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv-note: Deidentified free-text clinical notes. b.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Soo Yeon Kim, Saehoon Kim, Joongbum Cho, Young Suh Kim, In Suk Sol, Youngchul Sung, Inhyeok Cho, Minseop Park, Haerin Jang, Yoon Hee Kim, et al. A deep learning model for real-time mortality prediction in critically ill children. *Critical care*, 23(1):1–10, 2019.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Kwanhyung Lee, John Won, Heejung Hyun, Sangchul Hahn, Edward Choi, and Joohyung Lee. Self-supervised predictive coding and multimodal fusion advance patient deterioration prediction in fine-grained time resolution. *arXiv preprint arXiv:2210.16598*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. *arXiv preprint arXiv:2208.10240*, 2022.

Mengmeng Ma, Jian Ren, Long Zhao, S. Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence*, 2021.

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

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Neural Information Processing Systems*, 2021.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, 2022.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1, 2019.

Fred Qin, Vivek Madan, Ujjwal Ratan, Zohar Karnin, Vishaal Kapoor, Parminder Bhatia, and Taha Kass-Hout. Improving early sepsis prediction with multi modal learning. *arXiv preprint arXiv:2107.11094*, 2021.

Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, 5(1):149, 2022.

MinDong Sung, Sangchul Hahn, Chang Hoon Han, Jung Mo Lee, Jayoung Lee, Jinkyu Yoo, Jay Heo, Young Sam Kim, Kyung Soo Chung, et al. Event prediction model considering time and input error using electronic medical records in the intensive care unit: Retrospective study. *JMIR medical informatics*, 9(11):e26426, 2021.

Harini Suresh, Nathan Hunt, Alistair E. W. Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *ArXiv*, abs/1705.08498, 2017.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

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

Luís A Vale-Silva and Karl Rohr. Multisurv: Long-term cancer survival prediction using multimodal deep learning. *medRxiv*, pages 2020–08, 2020.

Miguel Vasco, Francisco S. Melo, and Ana Paiva. Mhvae: a human-inspired deep hierarchical generative model for multimodal representation learning. *ArXiv*, abs/2006.02991, 2020.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Yifan Wang and Ying Lan. Multi-view learning based on non-redundant fusion for icu patient mortality prediction. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1321–1325. IEEE, 2022.

Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K De Freitas, Ishan Paranjpe, Akhil Vaid, Jing Zhang, et al. Contrastive learning improves critical event prediction in covid-19 patients. *Patterns*, 2(12):100389, 2021.

Amir Zadeh, Minghai Chen, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine learning for healthcare conference*, pages 566–588. PMLR, 2020.

Appendix A. Detailed information on Experimental Setting

A.1. Cohort

A total of 53,150 patients with valid ICU admission and age over 18 were identified in the MIMIC-IV dataset. After excluding 20 patients due to missing vital sign records (i.e., pulse, systolic blood pressure, diastolic blood pressure, respiratory rate, and body temperature), the final cohort consists of 53,130 subjects. We further excluded patients using the process described in Section 4.1.1. Finally, for clinical event and intervention need prediction tasks, we followed the process illustrated in Section 4.2. In the end, we collect 42,813, 42,572, and 42,532 ICU subjects for mortality, vasopressor need, and intubation need prediction tasks respectively.

A.1.1. Clinical Objectives

In MIMIC-IV, we categorize Intubation (224385(which is item ID)), Intubation - Details (223059), Oral ETT (225307), Nasal ETT (225308), Unplanned Extubation (patient-initiated) (225468), Unplanned Extubation (non-patient initiated) (225477), Timeout Performed by (Intubation) (226188) as intubation task and Norepinephrine (221906), Dopamine (221662), Dobutamine (221289), Epinephrine (221289) as vasopressor task.

A.2. Data Sampler

As we implement real-time training incorporating up to 1440 hours per patient EHR data, there are an excessive amount of negatives. To address this imbalance during training, we employ a data sampler to impose equal proportions of positives and negatives during training.

A.3. EHR Time-Series Data Preparation

We select six vital-sign features, i.e., heart rate, respiration rate, diastolic and systolic blood pressure, temperature, and pulse oximetry. We gather ten features of laboratory result data, i.e., Hematocrit, Platelet, WBC, Bilirubin, pH, HCO3, Creatinine, Lactate, Potassium, and Sodium (Sung et al. (2021)).

A.4. EHR Text Data Preparation

A.4.1. Chief-complaint from MIMIC-ED

We use the chief complaint from MIMIC-ED as our EHR text data. When a chief complaint is available, it indicates that the subject has visited the Emergency Department (ED) prior to ICU admission.

A.4.2. Clinical Notes from MIMIC-IV-Note

We extract three sections of clinical notes: Chief Complaint, Past Medical History, and Medications on admission. We focus on these sections since we only require initial subject information to predict the occurrence of specific tasks within the hospital. To separate the extracted sections of clinical notes, we used [SEP] tokens. If the notes do not contain any of the three sections and if there were no available substitutes (such as Medical/Surgical History instead of Past Medical History), we only used the [SEP] token as the section separator. We exclude optional information about Past Medical History, such as Other Past Medical History, and also removed Medications-OTC from the Medications on Admission category, as it is not useful for predicting our tasks. The data table for the clinical note replacing chief complaint as our EHR text data is illustrated in Table 5, and the experimental results are illustrated in B.4. Note that the missing rate of the chief complaint is smaller than the clinical note (Table 1).

A.4.3. Pre-trained Text Embedder

We employed BioBERT, a pre-trained biomedical language representation model, to encode chief complaints or clinical notes from MIMIC-ED or MIMIC-IV-Note. The maximum

Table 5: Data statistics with patient numbers for mortality prediction, vasopressor need and intubation need prediction tasks with modality missing information. X-ray image is from MIMIC-CXR and clinical note text is from MIMIC-IV-Note.

(a) Mortality Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	3486	30870	430	3741	413	3873
Image Missing Rate	75.22%	77.84%	73.02%	78.56%	74.09%	78.23%
Text Missing Rate	10.01%	5.58%	6.28%	5.91%	8.23%	5.63%
(b) Vasopressor Need Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	9341	24822	1172	2969	1183	3085
Image Missing Rate	73.30%	79.16%	73.29%	79.69%	74.64%	79.00%
Text Missing Rate	6.22%	5.78%	7.0%	5.56%	5.75%	5.8%
(c) Intubation Need Prediction						
	Training		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
Patient Number	13450	20682	1665	2467	1716	2552
Image Missing Rate	72.71%	80.71%	73.51%	80.91%	73.66%	80.60%
Text Missing Rate	5.87%	5.94%	6.43%	5.63%	6.43%	5.63%

lengths were set to 128 and 512 for chief complaints and clinical notes, respectively. During our predictive training, BioBERT provided embeddings for both text data as a sequence of tokens. The pre-trained BioBERT utilized a vocabulary size of 28,996 to generate the sequence of embedding vectors.

A.5. EHR Image Data Preparation

A.5.1. Data Augmentations

To augment chest X-ray images, we applied a series of transformations during both the pre-training and fine-tuning phases. Specifically, we 1) resized each image to 256 $\times$ 256 pixels, 2) employed a set of random affine transformations, i.e., rotation, scaling, and translation, and 3) performed center-crop to obtain an image of size 224 $\times$ 224 pixels. For validation and testing, the images were resized to 256 $\times$ 256 pixels and underwent the same center crop operation to obtain an image of size 224 $\times$ 224 pixels. We consistently applied these procedures throughout our experiments.

A.5.2. Pre-trained Image Embedder

We selected our pre-trained SwinTransformer as an EHR image embedder (Liu et al. (2021)). We pre-trained Swin Transformer using CheXpert with the initial pre-trained weight from ImageNet-1K; we classify 14 binary radiology labels that were extracted from radiology reports via CheXpert(Irvin et al. (2019)). We optimized SwinTransformer using the binary cross entropy loss with learning rate sweep from 10^{-6} to 10^{-4}. To create pre-training

dataset, we extracted chest X-ray images of ICU patients from MIMIC-CXR and randomly split them by patient ID. Our training, validation, and test set consists of 213,016, 23,131, and 26,744 images, respectively. We employed the tiny version of the SwinTransformer, which uses 96 feature dimension, a patch size of 4×4, a window size of 7×7, and block depths of [2, 2, 6, 2].

Appendix B. Supplementary Results

B.1. Real-Time Time Embedding

In UMSE, we employ the time difference between the occurrence time and the current time, calculated as $t_{occurrence} - t_{current}$. Table 6 compares the predictive performance between using $t_{occurrence} - t_{current}$ and using $t_{occurrence}$ alone. Note that we calculated $t_{occurrence}$ as the time from the admission.

Table 6: **Effectiveness of Real-Time embedding** for three clinical event and intervention predictive tasks. The performance is measured in AUPRC and AUROC of validation dataset, averaged over 3 runs.

	Mortality		Vasopressor Need		Intubation Need	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
No Real-Time	0.828±0.005	0.961±0.003	0.834±0.001	0.926±0.001	0.855±0.003	0.928±0.001
Default (TSA)	**0.861±0.007**	**0.968±0.003**	**0.843±0.003**	**0.929±0.001**	**0.862±0.001**	**0.932±0.001**

B.2. Pre-training

We employed pretrained Swin Transformer fine-tuned on the Chexpert dataset and BioBERT pretrained with PubMed 1M [5]. In this section, we present the predictive performances when each pretrained encoder is substituted with an ImageNet pretrained Swin Transformer and BERT Tokenizer (see Table 7).

Table 7: **Result of Pretrained Encoder Effectiveness** for three clinical event and intervention predictive tasks. Here, INP refers to ImageNet pretrained Swin Transformer instead of Chexpert label pretrained Swin Transformer. BT refers to BERT tokenzier instead of BioBERT used in TSA.

	Mortality		Vasopressor Need		Intubation Need	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
TSA with INP	0.849±0.004	0.965±0.002	0.843±0.004	0.927±0.003	0.855±0.003	0.924±0.002
TSA with BT	0.845±0.003	0.961±0.002	0.839±0.003	0.926±0.003	**0.863±0.002**	**0.931±0.003**
Default (TSA)	**0.861±0.007**	**0.968±0.003**	**0.843±0.003**	**0.929±0.001**	0.862±0.001	0.932±0.001

5. https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

B.3. Multiple X-rays (EHR Image) per patient

UMSE can differentiate multiple non-time series data, therefore enabling to use multiple images/texts per patient. Since a single subject only possesses single text, we evaluated if using multiple X-ray images is beneficial. Specifically, we experimented TSA model training with up to the most recent three X-ray images whose results are summarized in Table 8. Note that for intubation need prediction task, the prediction performance slightly increases when using multiple images.

Table 8: Result of TSA model when maximum three of most recent X-ray images are used. The performance is measured in AUPRC and AUROC of validation dataset, averaged over 3 runs.

	Mortality		Vasopressor Need		Intubation Need	
L_{fusion}	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
TSA (Default)	**0.86±0.01**	**0.96±0.0**	**0.84±0.0**	**0.93±0.0**	0.86±0.0	0.93±0.0
TSA (maximum 3 images)	0.85±0.0	0.96±0.0	**0.84±0.0**	**0.93±0.0**	**0.87±0.0**	**0.94±0.0**

B.4. Clinical Note for EHR text

A recent release of the clinical note for the MIMIC-IV dataset enables us to compare the benefits of chief complaint versus clinical note, as detailed in Appendix A.4. Note the slight performance increase by using clinical notes instead of the chief complaint for both mortality and vasopressor use prediction (Table 9).

Table 9: Result of TSA model when MIMIC-IV-Note text data is utilized in the replacement of chief-complaint text from MIMIC-ED. The performance is measured in AUPRC and AUROC of validation dataset, averaged over 3 runs.

	Mortality		Vasopressor Need		Intubation Need	
L_{fusion}	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
With chief-complaint	0.861±0.007	0.968±0.003	0.843±0.003	0.929±0.001	**0.862±0.001**	**0.932±0.001**
With clinical note	**0.864±0.007**	**0.971±0.002**	**0.851±0.001**	**0.934±0.0**	0.855±0.002	0.928±0.001