
MULTIMODAL PROMPT LEARNING WITH IRREGULAR EHRs FOR ROBUST MONITORING OF CRITICAL CARE PATIENTS

Yixin Yang^{1,*} Yueyang Sun^{2,*} Weichen Liu³ Xianbing Zhao⁴ Sicen Liu^{1,†}

¹Faculty of Engineering, Shenzhen MSU-BIT University,

²International School, Beijing University of Posts and Telecommunications,

³School of Computer Science and Engineering, Southeast University,

⁴School of Artificial Intelligence and Computer Science, Jiangnan University

yyx@smbu.edu.cn, 2474931434@bupt.edu.cn, 213233033@seu.edu.cn,

zhaoxianbing_hitsz@163.com, liusc@smbu.edu.cn

ABSTRACT

Accurate assessment of patients in intensive care units (ICUs) is essential for timely clinical intervention and improved patient outcomes. Multimodal electronic health records (EHRs), including structured physiological time series and longitudinal clinical notes, provide complementary information for critical care prediction. However, in real-world clinical settings, individual modalities may be partially observed or entirely unavailable, resulting in substantial performance degradation for existing multimodal models. To address this challenge, we propose a multimodal prompt-learning framework for robust clinical prediction under diverse missing-modality scenarios. The proposed framework introduces four complementary types of prompts: generative prompts, missing-signal prompts, missing-type prompts, and temporal prompts. Generative prompts construct surrogate latent representations for unavailable modalities, while missing-signal prompts distinguish observed representations from generated ones. Missing-type prompts condition the model on different modality-availability configurations, whereas temporal prompts perform condition-specific aggregation over temporally encoded clinical sequences. Together, these prompts enable the model to capture missingness-aware intramodal dependencies and cross-modal interactions within a unified architecture. Extensive experiments demonstrate that our method outperforms existing approaches across evaluation metrics on two missingness settings. Ablation and robustness analyses further verify the complementary contributions of the four prompt types and the effectiveness of the proposed framework for clinical prediction from incomplete multimodal EHR data.

1 Introduction

Intensive care units (ICUs) provide continuous care for patients with life-threatening conditions, such as severe trauma [1], sepsis [2], and organ failure [3]. The early hours following ICU admission are particularly critical, as patients may experience rapid physiological deterioration and are highly vulnerable to delayed or inappropriate clinical decisions [4, 5]. Accurate prediction of in-hospital mortality based on patients’ clinical information collected during the first 48 hours of ICU admission is therefore essential for timely risk assessment, resource allocation, and clinical intervention. The widespread adoption of electronic health records (EHRs) has substantially transformed the documentation and analysis of patients’ health conditions in ICUs [6, 7, 8]. EHRs contain heterogeneous and complementary sources of clinical information, including multivariate irregularly sampled time series (MISTS), such as laboratory measurements and vital signs, as well as longitudinal clinical notes documenting patients’ symptoms, diagnoses, treatments, and disease progression. The increasing availability of these data has facilitated the development of deep learning models for various clinical applications [9, 10, 11], including mortality prediction [12], sepsis onset prediction [13], phenotype classification [14], and medication recommendation [15, 16].

*Equal Contribution

†Corresponding Author

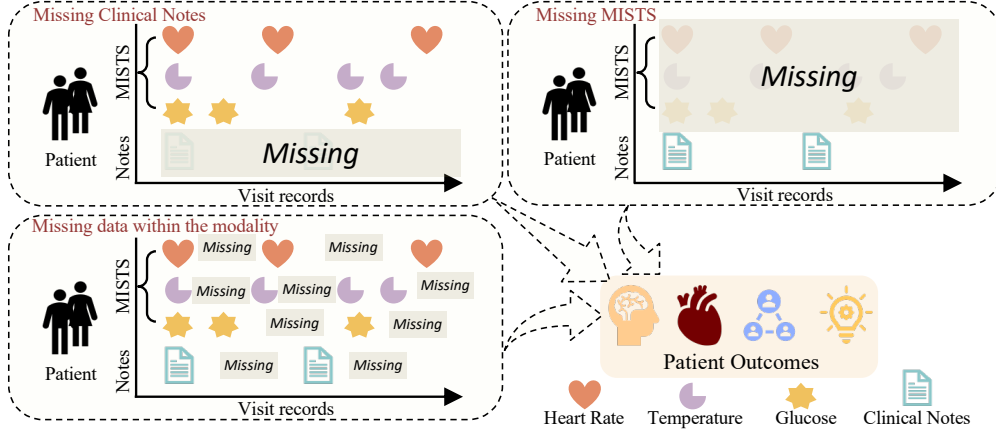


Figure 1: Examples of different types of missing information in EHR data include missing clinical notes, missing MISTS, and missing data within the modality. The heart rate and temperature are monitored regularly with different frequencies, and glucose is a laboratory test ordered at irregular time intervals. The clinical notes are free text.

Existing studies have developed multimodal architectures to integrate irregular physiological measurements and clinical notes [17, 18]. However, both modalities may not always be available during inference, as real-world EHRs may contain incomplete data or have one modality entirely missing due to equipment failures, delayed documentation, data corruption, privacy restrictions, or variations in clinical workflows. As illustrated in Figure 1, different missing-modality configurations introduce distinct information gaps and may substantially degrade mortality prediction. Moreover, observations within each modality are temporally irregular: physiological variables are recorded at uneven intervals, while clinical notes are documented at different stages of an ICU stay. Therefore, a robust model should distinguish observed representations from generated ones, adapt its fusion strategy to different modality-availability conditions, and preserve the irregular temporal information encoded in the available clinical streams.

Prompt learning provides a flexible mechanism for adapting model representations to different input conditions through learnable parameters [19, 20]. Motivated by this property, we propose a multimodal prompt-learning framework for in-hospital mortality prediction using EHR data collected during the first 48 hours of ICU admission. The framework incorporates four complementary prompts. Generative prompts provide surrogate representations for unavailable modalities, missing-signal prompts distinguish observed features from surrogate representations, missing-type prompts characterize modality-availability configurations, and temporal prompts perform condition-specific aggregation over the irregular temporal information. These prompts are jointly integrated into a multimodal Transformer to learn missingness-aware temporal, intramodal, and cross-modal representations from incomplete EHR data.

We conduct extensive experiments on two large publicly available benchmark datasets, MIMIC-III [21] and MIMIC-IV [22], validating the effectiveness and superiority of our proposed method.

The main contributions of this work are three-fold:

- We present a novel framework via prompt learning for monitoring critical care patients that is capable of handling missing modalities efficiently with irregular EHR.
- We propose four types of prompts to tackle the problem of missing modalities in irregular EHR data. These prompts can generate missing information, learn irregular temporal patterns, and effectively capture both intra- and inter-modality information.
- Extensive experiments on large public available benchmark dataset MIMIC-III and MIMIC-IV dataset demonstrate that our proposed approach outperforms the compared methods.

2 Related Work

Multimodal Modeling of Irregular EHRs. Electronic health records commonly contain structured multivariate irregular time series (MISTS) and temporally irregular clinical notes. Earlier studies address irregularity through time-aware decay, learnable interpolation, set-based representations, or attention over continuous-time embeddings [23, 24, 25, 26]. Recent research has increasingly focused on jointly modeling structured and unstructured EHR modalities. Zhang et al. propose a unified irregular-EHR framework that models MISTS and clinical-note sequences separately and

integrates them through temporally interleaved attention [17]. Liu and Chen further incorporate domain and memory knowledge into irregular multimodal EHR representations [18], while CTPD discovers and aligns cross-modal temporal patterns shared across patients [27]. These approaches improve temporal representation and multimodal fusion, but primarily focus on extracting information from the available input streams.

Learning with Missing Clinical Modalities. Missing modalities are common in clinical data because measurements, notes, images, and reports are collected under heterogeneous clinical workflows. Earlier work explored missing-modality learning through cross-modal translation, modality imagination, and missing-modality-aware prompting [28, 29, 30]. Recent clinical studies consider more realistic settings: MUSE jointly addresses missing modalities and labels [31]. Wang et al. address missing chest radiographs, radiology reports, or structured clinical variables [32], while Liang et al. model missing patterns related to patient conditions or clinical decisions [33]. More recently, Zhao et al. combine modality reconstruction with prompt-guided adaptation to handle missing MRI sequences during both training and inference [34]. Despite these advances, existing methods primarily emphasize latent compensation, representation consistency, or robust aggregation. Our work instead couples prompt-conditioned cross-modal generation with explicit representations of latent-stream provenances, whole-modality availability, and irregular temporal progression for paired MISTS and longitudinal clinical notes.

Prompt Learning for Missing Modalities. Prompt-based adaptation uses task-specific natural-language templates in NLP and learnable multimodal prompt tokens in vision-language models [19, 20]. Lee et al. introduces multimodal prompts for visual recognition with missing inputs [30]. Most closely related to our work, Guo et al. propose generative, missing-signal, and missing-type prompts for multimodal sentiment analysis and emotion recognition under missing modalities [35]. However, these methods are designed for general multimodal settings rather than EHR data, and none of them directly addresses missing modalities in irregular EHR with heterogeneous temporal patterns. We adapt this prompt-based missing-modality paradigm to irregular clinical data by introducing temporal prompts and integrating prompt-conditioned reconstruction with time-aware MISTS and clinical-note encoders for ICU outcome prediction.

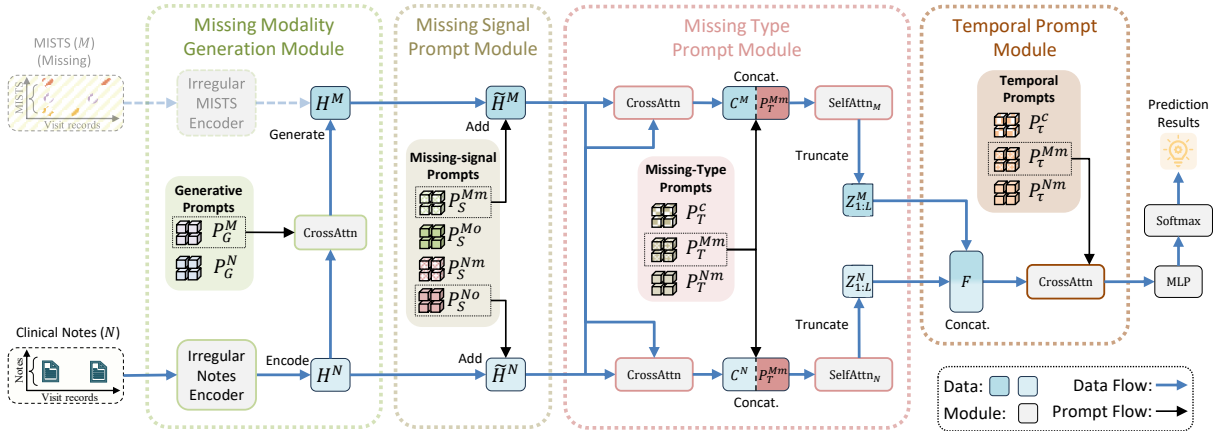


Figure 2: Overview of the proposed prompt-based framework for robust outcome prediction from incomplete irregular multimodal EHRs, based on a Notes-only example. Given a notes-only input condition, irregular notes encoders extract the available representation, and the Missing Modality Generation Module reconstructs the absent MISTS stream. Missing-signal prompts augment each token with a provenance embedding indicating whether its modality stream is observed or prompt-generated, while missing-type prompts inject the sample-level modality-availability condition into both post-fusion streams. The resulting representations are fused and summarized by temporal prompts for patient outcome prediction.

3 Method

As illustrated in Figure 2, the proposed framework takes multivariate irregularly sampled time series and longitudinal clinical notes as input. Four types of learnable prompts are incorporated into a multimodal Transformer to model irregular temporal information and diverse missing-modality patterns. The resulting multimodal representation is subsequently fed into a prediction head to estimate the probability of in-hospital mortality.

3.1 Problem Formulation

We formulate critical care monitoring as a patient-level prediction problem from irregular multimodal EHRs. Each sample from an ICU stay is observed within an early clinical window, that is, the first 48 hours after admission. The sample index is omitted for clarity. The EHR data comprise two heterogeneous modalities: $v \in \{M, N\}$, where M denotes MISTS and N denotes the clinical notes. Let $m \in \{c, Mm, Nm\}$ denote the sample-level modality condition, where $m = c$ indicates that both modalities are observed, $m = Mm$ indicates MISTS is missing, and $m = Nm$ indicates notes is missing. The input can be formulated as

$$X = \{\{(t_i^M, x_i^M)\}_{i=1}^{T_M}, \{(t_j^N, x_j^N)\}_{j=1}^{T_N}, m\} \quad (1)$$

with observation time t_k^v , multivariate clinical value x_k^v , and number of observed events T_v . For a missing modality, the corresponding input sequence is unavailable. The objective is to learn $\hat{y} = f_\theta(X) \in [0, 1]$, the predicted risk of 48-hour in-hospital mortality.

3.2 Missing Modality Generation Module

Since MISTS and clinical notes differ in their structures and temporal characteristics, we employ separate modality-specific encoders to obtain their latent representations when they are available: the UTDE encoder [17] for MISTS and the Clinical-Longformer-based mTAND_{txt} encoder [36, 17] for clinical notes.

$$H^M = \text{Enc}_M(X^M, T^M) \quad (2)$$

$$H^N = \text{Enc}_N(X^N, T^N) \quad (3)$$

where X^M, T^M denote the numerical observations, timestamps of MISTS, respectively, while X^N and T^N denote the note embeddings and their timestamps.

When one modality is missing, we employ the Missing Modality Generation Module to construct a surrogate representation for the unavailable modality. As depicted in Figure 2, when clinical notes are observed and MISTS are missing ($m = Mm$), H^N is obtained directly from the irregular notes encoder, while H^M is generated from H^N using the generative prompt P_G^M through cross-attention [37]:

$$H^M = \text{CrossAttn}(P_G^M, H^N, H^N). \quad (4)$$

Here, $P_G^M \in \mathbb{R}^{L \times d}$ is a learnable generative prompt, where L denotes the number of positions on the unified temporal grid and d denotes the latent dimension shared by the two modalities. P_G^M serves as the query, while the observed clinical-note representation H^N serves as both the keys and values. The resulting H^M is a time-aligned surrogate representation for the missing MISTS modality.

3.3 Missing Signal Prompt Module

Although the generated representation H^M restores the missing MISTS stream, it should not be treated in the same way as the directly observed clinical-note representation H^N . We therefore introduce the Missing Signal Prompt Module to indicate whether each modality representation is observed or generated.

The module annotates each stream with its provenance using four learnable prompts for both modalities: $P_S^{Mm}, P_S^{Mo}, P_S^{Nm}, P_S^{No} \in \mathbb{R}^{L \times d}$, indicating that the modality is missing (Mm, Nm) or observed (Mo, No). Therefore, the corresponding prompts are added as follows:

$$\tilde{H}^M = H^M + P_S^{Mm}, \quad (5)$$

$$\tilde{H}^N = H^N + P_S^{No}. \quad (6)$$

Here, by adding P_S^{Mm} and P_S^{No} to the generated MISTS representation and observed clinical-note representation, respectively, each time-aligned token now carries a position-specific marker of its source, so that the downstream fusion is able to weigh the observed and synthesized features asymmetrically rather than blending them uniformly.

3.4 Missing Type Prompt Module

Although the pair of provenance prompts implicitly identifies the modality condition, this information remains distributed across the two streams. We therefore introduce a shared missing-type prompt that provides both post-fusion branches with an explicit global availability condition.

Specifically, the Missing Type Prompt Module provides an explicit global condition through three learnable missing-type prompts: $P_T^c, P_T^{Mm}, P_T^{Nm} \in \mathbb{R}^{L_p \times d}$, where L_p is the number of type tokens. The same missing-type prompt is used for both modality streams to provide a global condition.

The two modality representations first exchange information through bidirectional cross-attention:

$$C^M = \text{CrossAttn}(\tilde{H}^M, \tilde{H}^N, \tilde{H}^N), \quad (7)$$

$$C^N = \text{CrossAttn}(\tilde{H}^N, \tilde{H}^M, \tilde{H}^M). \quad (8)$$

When modality M is missing, P_T^{Mm} is selected and appended to both cross-attended representations:

$$Z^M = \text{SelfAttn}_M([C^M; P_T^{Mm}]), \quad (9)$$

$$Z^N = \text{SelfAttn}_N([C^N; P_T^{Mm}]). \quad (10)$$

Here, $[\cdot; \cdot]$ denotes sequence-wise concatenation and $Z^M, Z^N \in \mathbb{R}^{(L+L_p) \times d}$. The two self-attention stacks share the same architecture but have separate parameters. By using P_T^{Mm} in both streams, the fusion module is explicitly informed that the MISTS representation is generated while the clinical-note representation is observed.

3.5 Temporal Prompt Module

The final prediction still requires a patient-level summary over the irregular clinical timeline. A fixed pooling operation cannot adaptively identify the temporal positions that are most relevant under different modality-availability conditions. We therefore introduce the Temporal Prompt Module for condition-specific temporal aggregation.

Because Z^M and Z^N contain both temporal tokens and the appended missing-type prompt tokens, we retain their first L temporal tokens and concatenate them along the feature dimension:

$$F = \text{Concat}_{\text{feat}}(Z_{1:L}^M, Z_{1:L}^N) \in \mathbb{R}^{L \times 2d}. \quad (11)$$

Here, $Z_{1:L}^M$ and $Z_{1:L}^N$ denote the first L time-aligned modality tokens in Z^M and Z^N . The remaining L_p tokens correspond to the appended missing-type prompts and are excluded from temporal aggregation. $\text{Concat}_{\text{feat}}$ denotes concatenation along the feature dimension.

To adapt temporal aggregation to different modality-availability conditions, we maintain three learnable temporal prompts: $P_\tau^c, P_\tau^{Mm}, P_\tau^{Nm} \in \mathbb{R}^{1 \times 2d}$, corresponding to three modality missing condition. When modality M is missing, P_τ^{Mm} is selected:

$$s_{Mm} = \text{CrossAttn}(P_\tau^{Mm}, F, F). \quad (12)$$

Here, P_τ^{Mm} serves as the query, while F provides the keys and values. The resulting representation $s_{Mm} \in \mathbb{R}^{1 \times 2d}$ summarizes the temporal information most relevant to the Notes-only condition.

Finally, the summarized representation is processed by a residual two-layer MLP and a linear classifier:

$$r_{Mm} = s_{Mm} + \text{MLP}(s_{Mm}), \quad (13)$$

$$\hat{y} = [\text{Softmax}(W_o r_{Mm} + b_o)]_{\text{mortality}} \quad (14)$$

Here, $r_{Mm} \in \mathbb{R}^{2d}$ denotes the resulting patient-level representation after residual connection. $W_o \in \mathbb{R}^{2 \times 2d}$ and $b_o \in \mathbb{R}^2$ are the learnable weight matrix and bias vector of the binary classifier, respectively. $\hat{y} \in [0, 1]$ denotes the predicted probability of in-hospital mortality.

4 Experiment

4.1 Datasets

We evaluate our method on multimodal ICU cohorts from MIMIC-III [21] and MIMIC-IV v3.1.¹ For MIMIC-III, we follow the processed irregular-EHR benchmark cohort of Zhang et al. [17]; for MIMIC-IV, clinical notes are taken from MIMIC-IV-Note.² The task is 48-hour in-hospital mortality prediction (48-IHM), where each sample uses the first 48 hours after ICU admission to predict hospital mortality. It is a binary classification problem with label imbalance with death to discharge ratio of approximately 1:7. ICU stays shorter than 48 hours are excluded.

Each sample contains a structured MISTS modality and a clinical-note modality. MISTS includes 17 benchmark clinical variables with irregular timestamps and observation masks; we also build a one-hour value-and-mask representation and normalize continuous variables using training-set statistics. Clinical notes are aligned by ICU admission time, and at most the latest five notes within the 48-hour window are tokenized with Clinical-Longformer and truncated or padded to 1024 tokens per note.

4.2 Evaluation Metrics

Following prior work [17], we evaluate 48-IHM as an imbalanced binary prediction task in which identifying mortality cases is clinically important. We report F1 to characterize classification performance at a fixed decision threshold, AUPRC to evaluate positive-case retrieval under label imbalance, and AUROC to provide a complementary measure of overall discrimination. Because the dataset is class-imbalanced and mortality is the minority class, we use AUPRC as the primary metric for model selection, as it better reflects the model’s ability to identify mortality cases.

4.3 Baselines

To ensure a comprehensive comparison, we evaluate our method against three groups of baselines: MISTS-only approaches, Notes-only approaches, and Full-modality approaches. For MIMIC-III, we follow the published irregular-EHR baseline results [17]; for MIMIC-IV, we evaluate the listed baselines under our data split.

For the *MISTS-only* setting on MIMIC-III, we follow the benchmark comparison with Imputation [38], IP-Net [24], mTAND [26], GRU-D [23], RAINDROP [39], and UTDE [17]. On MIMIC-IV, we evaluate the regular time-series backbones IP-Net, GRU-D, SeFT [25], DGM²-O [40], mTAND, RAINDROP, MTGNN [41], and UTDE.

For the *Notes-only* setting on both datasets, we compare with the note-based models actually used in our experiments: Flat [42], HierTrans [43], T-LSTM [44], FT-LSTM [45], GRU-D [23], and mTAND-txt [17].

For the *Full-modality* setting on MIMIC-IV, we compare with representative multimodal EHR and feature-fusion baselines, including MMIMIC [17], CTPD [27], MedFuse [46], DAFT [47], MMTM [48], and DrFuse [49]. Under each missing-modality setting, our model receives only the available modality and reconstructs the missing stream through prompt-based cross-modal generation.

4.4 Implementation Details

Architecture and Training Details. We use a shared embedding dimension of $d = 128$ and a unified timeline of length $T = 48$. Clinical notes are pre-tokenized by Clinical-Longformer with a maximum token length of 1024, and are temporally aligned to the unified timeline using mTAND-txt. For the MISTS stream, we adopt UTDE, which incorporates a learnable gate based on mTAND-ts. The text encoder is initialized from a pretrained Clinical-Longformer checkpoint, while the remaining task-specific modules are trained from scratch. Each run is trained up to a fixed maximum number of epochs, and model selection uses validation AUPRC under the corresponding missing-modality scenario.

Prompt initialization. Generative prompt queries are initialized from the training-set global means of the projected target-modality embeddings. The modality-signal and missing-type prompts are initialized to zero, whereas the missing-mode-specific temporal prompts are independently initialized from a standard normal distribution. All prompt parameters are jointly optimized with the remaining task-specific modules.

¹<https://physionet.org/content/mimiciv/3.1/>

²<https://physionet.org/content/mimic-iv-note/2.2/>

Training-Time Modality Dropout. To expose the model to incomplete inputs during training, we apply modality-level dropout to the paired multimodal training samples. For each complete training batch, we sample a Bernoulli variable with probability η . If dropout is not activated, both MISTS and clinical notes are kept and the sample is trained under the complete case $m = c$. If dropout is activated, one modality is removed at the input level according to the target missing scenario.

4.5 Main Results

We first examine our method’s robustness when either modality is unavailable, followed by assessing whether the proposed framework retains competitive performance when both modalities are observed.

Table 1: Main results on MIMIC-III under single-modality settings. Best results are shown in bold and second-best results are underlined.

Metric	<i>MISTS-only setting</i>							<i>Notes-only setting</i>						
	Imputation	IP-Net	mTAND	GRU-D	RAINDROP	UTDE	Ours	Flat	HierTrans	T-LSTM	FT-LSTM	GRU-D	mTAND-txt	Ours
F1	0.3973	0.3722	0.4387	0.4282	0.3946	<u>0.4526</u>	0.4961	0.3978	0.4876	0.5032	0.4851	0.5101	<u>0.5257</u>	0.5284
AUPRC	0.4436	0.3936	0.4754	0.4590	0.3623	<u>0.4964</u>	0.5028	0.5169	0.5298	0.5257	0.5439	0.5434	<u>0.5605</u>	0.5853

Table 2: Main results on MIMIC-IV under single-modality settings. Best results within each setting and metric are shown in bold and second-best results are underlined.

<i>MISTS-only setting</i>									
Metric	GRU-D	IP-Net	SeFT	MTGNN	DGM ² -O	mTAND	RAINDROP	UTDE	Ours
AUROC	0.8293	0.8580	0.6536	0.8314	0.8394	0.8540	0.8323	0.8620	0.8586
AUPRC	0.4541	<u>0.5154</u>	0.2369	0.4826	0.4832	0.5025	0.4447	0.5099	0.5202
F1	0.3114	0.4044	0.0672	0.4191	0.4190	0.4168	0.3219	0.4856	<u>0.4585</u>

<i>Notes-only setting</i>							
Metric	T-LSTM	GRU-D	HierTrans	FT-LSTM	Flat	mTAND-txt	Ours
AUROC	0.6748	0.8120	<u>0.8148</u>	0.8109	0.8112	0.8099	0.8181
AUPRC	0.2677	0.4329	<u>0.4380</u>	<u>0.4384</u>	0.4245	0.4170	0.4509
F1	0.1394	<u>0.4121</u>	0.3798	0.3937	0.4081	0.2229	0.4643

Missing-modality Comparison To examine whether the proposed framework can recover useful information beyond that captured by unimodal models, Tables 1 and 2 compare our method with strong MISTS-only and Notes-only baselines on two independent cohorts. Our method achieves the highest AUPRC in all four settings. Since mortality cases are underrepresented and AUPRC is our primary metric, this consistent result provides strong evidence that our prompt-based framework improves the identification and ranking of high-risk patients under missing modalities. Moreover, the gains are more consistent across metrics in the Notes-only setting than in the MISTS-only setting. This asymmetry suggests that reconstructing MISTS from clinical notes is generally more stable than reconstructing clinical notes from MISTS.

Table 3: Full-modality results on MIMIC-IV, where both MISTS and clinical notes are available at test time. Best results are shown in bold and second-best results are underlined.

Metric	MMTM	DAFT	MedFuse	DrFuse	CTPD	MMIMIC	Ours
AUROC	0.8776	0.8488	0.8523	0.8588	0.8694	0.8826	<u>0.8785</u>
AUPRC	<u>0.5761</u>	0.4910	0.5041	0.5141	0.5384	0.5767	0.5587
F1	0.4823	0.2541	0.3614	0.4698	0.4060	<u>0.4839</u>	0.5285

Full-modality Comparison. We further evaluate whether our method maintains its performance when both modalities are available. As shown in Table 3, our method achieves the best F1 and the second-best AUROC. Compared with MMIMIC, which serves as our backbone, it substantially improves F1, with only modest decreases in AUROC and AUPRC. These results indicate that the proposed method largely preserves the backbone’s predictive capability under complete observations in addition to handling missing modalities effectively.

4.6 Ablation Study

To better understand the contribution of each design choice, we conduct ablation studies from two perspectives: contribution of different prompts and training-time modality dropout rate.

Table 4: Prompt ablation under the Notes-only setting on MIMIC-III, including individual, three-prompt, and cumulative configurations. A check mark indicates that the corresponding prompt is enabled. Best and second-best results are shown in bold and underlined, respectively.

Prompt Configuration				Performance	
P_G	P_S	P_T	P_τ	AUPRC	AUROC
				0.5383	0.8759
✓				0.5566	0.8841
	✓			0.5572	0.8839
		✓		0.5527	0.8822
			✓	0.5412	0.8782
✓	✓			0.5695	0.8864
✓	✓	✓		0.5768	0.8917
✓	✓	✓	✓	0.5853	<u>0.8904</u>
	✓	✓	✓	0.5597	0.8874
✓		✓	✓	0.5700	0.8872
✓	✓		✓	0.5687	0.8875

Prompt Ablation. Table 4 demonstrates the effectiveness and complementarity of the four prompts. Enabling any single prompt yields modest improvements in both AUPRC and AUROC, indicating that each provides useful task-specific information. The cumulative configurations show a clear progression: P_G reconstructs the missing modality, P_S identifies the provenance of generated and observed features, P_T adapts cross-modal interaction to the modality-availability pattern, and P_τ performs condition-specific temporal aggregation. As these functions are progressively integrated, AUPRC increases steadily and reaches its highest value of 0.5853 with all prompts enabled. Consistently, removing any prompt from the complete configuration reduces AUPRC to varying degrees, supporting that the four prompts make non-redundant contributions and work together to improve missing-modality prediction.

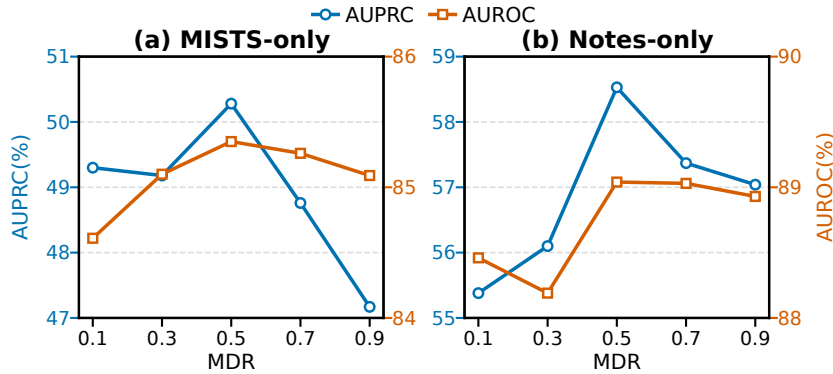


Figure 3: Modality dropout rate ablation on MIMIC-III. Each panel reports AUPRC on the left y-axis and AUROC on the right y-axis for $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Dropout Rate. We vary the training-time modality dropout rate η to examine how exposure to simulated missing inputs affects robustness. As shown in Figure 3, both settings show a non-monotonic trend. When the dropout rate is too low, the model is not sufficiently exposed to incomplete cases during training; when it is too high, effective observed supervision is reduced and the generated representations become less reliable. These observations suggest that a moderate-to-high dropout rate forces the model to learn stronger missing-modality compensation and more robust cross-modal interaction under incomplete observations.

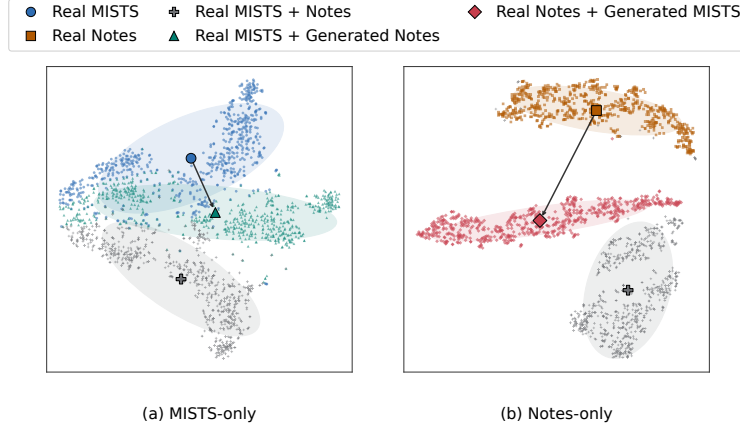


Figure 4: t-SNE visualization of classifier-level representations on MIMIC-III. Panel (a) and (b) show the MISTS-only and Notes-only settings respectively. Each panel compares the observed single-modality representation, its prompt-completed representation, and the fully observed MISTS+Notes reference.

Representation Visualization. To examine whether prompt completion recovers meaningful cross-modal structure, we visualize the learned representations on MIMIC-III. As shown in Figure 4, the completed representations shift from the single-modality region toward the Full-modality region while retaining the structure of the observed modality. This controlled shift indicates that prompt completion supplements rather than overwrites the available evidence, reducing the representation gap between incomplete and complete EHR inputs.

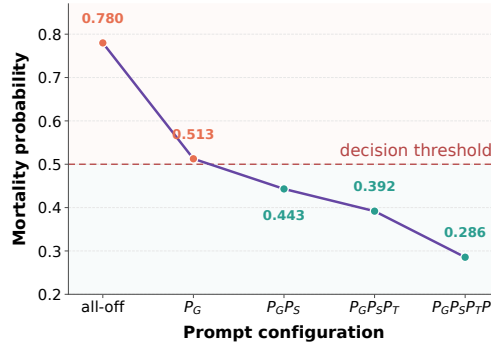


Figure 5: Evolution of the predicted mortality risk for a representative patient in the MIMIC-III Notes-only setting as the prompt components are cumulatively activated. *All-off* indicates that no prompts are used. Predictions above the decision threshold are classified as mortality, whereas those below the threshold are classified as survival.

4.7 Case Study

To examine how the prompts affect individual predictions and how the model integrates evidence across clinical notes, we conduct a case study on a representative patient. Figure 5 illustrates how the prompts correct a survival case under the MIMIC-III Notes-only setting. Without prompts, the model incorrectly predicts mortality. P_G produces the largest risk reduction, showing that the reconstructed MISTS stream supplies the main complementary evidence. P_S then moves the prediction below the decision threshold, while P_T and P_r maintain the correct decision. This progression indicates that reconstruction drives the correction, whereas the remaining prompts refine how generated and observed evidence is interpreted.

Figure 6 further shows that the prediction is based on evidence distributed across multiple notes. The impaired responsiveness in N2 and severe hypoglycemia in N4 support mortality, whereas N1 and N5 contain evidence favoring survival. The model therefore reaches the correct prediction by balancing conflicting clinical observations rather than relying on a single note.

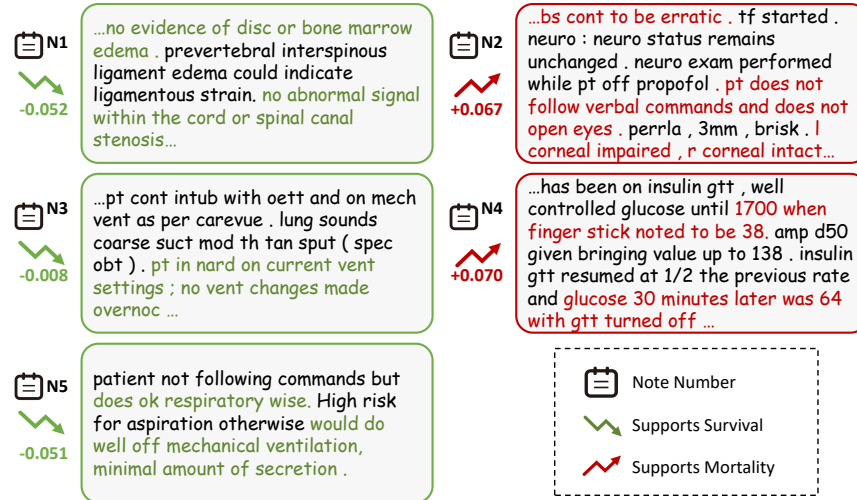


Figure 6: Changes in predicted mortality risk after removing each note for the same patient. Negative scores shown in green indicate evidence supporting survival, whereas positive scores shown in red indicate evidence supporting mortality.

5 Conclusion

In this paper, we present a unified multimodal prompt-learning framework for robust mortality prediction from irregular and incomplete EHRs. We proposed generative prompts, missing-signal prompts, missing-type prompts, and temporal prompts to capture information under missing-modality conditions. Ablation study and case study further confirm that the four prompts play complementary roles in aligning incomplete inputs with the full-modality decision space. Future work can evaluate the framework on broader clinical prediction tasks and datasets, and extend it to settings involving more modalities and complex real-world missingness patterns.

References

- [1] Samuel A Tisherman and Deborah M Stein. Icu management of trauma patients. *Critical Care Medicine*, 46(12):1991–1997, 2018.
- [2] Corinne Alberti, Christian Brun-Buisson, Hilmar Burchardi, Claudio Martin, Sergey Goodman, Antonio Artigas, Alberto Sicignano, Mark Palazzo, Rui Moreno, Ronan Boulmé, et al. Epidemiology of sepsis and infection in icu patients from an international multicentre cohort study. *Intensive care medicine*, 28:108–121, 2002.
- [3] Bekele Afessa, Ognjen Gajic, and Mark T Keegan. Severity of illness and organ failure assessment in adult intensive care units. *Critical care clinics*, 23(3):639–658, 2007.
- [4] David J Cullen, Bobbie Jean Sweitzer, David W Bates, Elisabeth Burdick, Amy Edmondson, and Lucian L Leape. Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units. *Critical care medicine*, 25(8):1289–1297, 1997.
- [5] María José Otero-López, Pablo Alonso-Hernández, José Angel Maderuelo-Fernández, Beatriz Garrido-Corro, Alfonso Domínguez-Gil, and Angel Sánchez-Rodríguez. Preventable adverse drug events in hospitalized patients. *Medicina clinica*, 126(3):81–87, 2006.
- [6] Julia Adler-Milstein, Catherine M DesRoches, Peter Kralovec, Gregory Foster, Chantal Worzala, Dustin Charles, Talisha Searcy, and Ashish K Jha. Electronic health record adoption in us hospitals: progress continues, but challenges persist. *Health affairs*, 34(12):2174–2180, 2015.
- [7] Hilal Atasoy, Brad N Greenwood, and Jeffrey Scott McCullough. The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. *Annual review of public health*, 40:487–500, 2019.
- [8] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10, 2018.

- [9] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.
- [10] Shams Forruque Ahmed, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M Mofijur, ABM Shawkat Ali, and Amir H Gandomi. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617, 2023.
- [11] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.
- [12] Feng Xie, Bibhas Chakraborty, Marcus Eng Hock Ong, Benjamin Alan Goldstein, Nan Liu, et al. Autoscore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10):e21798, 2020.
- [13] John Karlsson Valik, Logan Ward, Hideyuki Tanushi, Anders F Johansson, Anna Färnert, Mads Lause Mogensen, Brian W Pickering, Vitaly Herasevich, Hercules Dalianis, Aron Henriksson, et al. Predicting sepsis onset using a machine learned causal probabilistic network algorithm based on electronic health records data. *Scientific reports*, 13(1):11760, 2023.
- [14] Tetiana Hourani, Alexis Perez-Gonzalez, Khashayar Khoshmanesh, Rodney Luwor, Adrian A Achuthan, Sara Baratchi, Neil M O’Brien-Simpson, and Akram Al-Hourani. Label-free macrophage phenotype classification using machine learning methods. *Scientific Reports*, 13(1):5202, 2023.
- [15] Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems*, 41(1):1–24, 2023.
- [16] Sicen Liu, Xiaolong Wang, Jingcheng Du, Yongshuai Hou, Xianbing Zhao, Hui Xu, Hui Wang, Yang Xiang, and Buzhou Tang. Shape: A sample-adaptive hierarchical prediction network for medication recommendation. *IEEE Journal of Biomedical and Health Informatics*, 27(12):6018–6028, 2023.
- [17] Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pages 41300–41313. PMLR, 2023.
- [18] Sicen Liu and Hao Chen. Knowledge injected multimodal irregular ehrrs model for medical prediction. In *International Workshop on Trustworthy Artificial Intelligence for Healthcare*, pages 25–39. Springer, 2024.
- [19] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021.
- [20] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.
- [21] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- [22] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [23] 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.
- [24] Satya Narayan Shukla and Benjamin M. Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2019.
- [25] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4353–4363, 2020.
- [26] Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series, 2021.
- [27] Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. CTPD: Cross-modal temporal pattern discovery for enhanced multimodal electronic health records analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6783–6799, 2025.

- [28] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899, 2019.
- [29] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021.
- [30] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.
- [31] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Michael Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, volume 2024, pages 55538–55557, 2024.
- [32] Muyu Wang, Shiyu Fan, Yichen Li, Zhongrang Xie, and Hui Chen. Missing-modality enabled multi-modal fusion architecture for medical data. *Journal of Biomedical Informatics*, 164:104796, 2025.
- [33] Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. Causal representation learning from multimodal clinical records under non-random modality missingness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28779–28796, 2025.
- [34] Yunpeng Zhao, Cheng Chen, Qing You Pang, Yibing Fu, Quanzheng Li, Carol Tang, Beng Ti Ang, and Yueming Jin. Tackling dual-stage missing modalities in brain tumor segmentation via robust modality reconstruction and prompt-guided modality adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13314–13322, 2026.
- [35] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, 2024.
- [36] Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-Longformer and clinical-BigBird: Transformers for long clinical sequences. *Journal of Biomedical Informatics*, 130:104069, 2022.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [38] Zachary C Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *Machine learning for healthcare conference*, pages 253–270. PMLR, 2016.
- [39] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations (ICLR)*, 2022.
- [40] Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan B Davidson. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 651–659, 2021.
- [41] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 753–763, 2020.
- [42] Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, 2021.
- [43] Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844, 2019.
- [44] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 65–74, 2017.
- [45] 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.

- [46] Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 479–503. PMLR, 2022.
- [47] Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3d image and tabular data via the dynamic affine feature map transform. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12905 of *Lecture Notes in Computer Science*, pages 688–698. Springer, 2021.
- [48] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020.
- [49] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16416–16424, 2024.