

When Clinical Notes Are Incomplete: Fusing Clinical Notes and Time Series Using Diffusion-based Models

Feilong Wu, Kejing Yin, and William K. Cheung

Department of Computer Science, Hong Kong Baptist University

{csflwu, cskjyin, william}@comp.hkbu.edu.hk

Abstract

Fusing charted times series with clinical notes has been found useful for predictive analytics. However, the incompleteness of clinical notes can undermine prediction quality. We thus propose a novel conditional diffusion-based model called CODE to recover the missing semantics of clinical notes, which integrates EHR time series into diffusion. CODE learns to align a diffusion process with unknown levels of note incompleteness using contrastive learning, by which missing semantics are recovered via the denoising process. We evaluate the effectiveness of CODE by conducting experiments on real-world EHR data for three prediction tasks. Empirical results show that CODE with the capability to recover the missing semantics can outperform state-of-the-art baselines.

1 Introduction

Recent advances leverage multimodal temporal EHR data, particularly clinical notes and time series of clinical measurements, to model patient conditions and forecast treatment needs (Zhang et al., 2023b; Cui et al., 2024; Lyu et al., 2023; Han et al., 2024; Yun et al., 2025; Zhu et al., 2024; King et al., 2023; Liu and Chen, 2024). Time series data and clinical notes are complementary in providing a comprehensive view of a patient’s condition. The time series often include vital signs and lab test results, which are typically acquired using equipment, while clinical notes (e.g., nursing notes) are written by professionals to report the disease progression observed during nursing activities. Work experience, competence, and education can be factors that affect nurses’ interpretations of disease progression (Cohan et al., 2016; Charalambous and Goldberg, 2016), in which case some critical information can be omitted. Consequently, performance declines in predictive analytics. This incompleteness issue is challenging as missingness is hard to

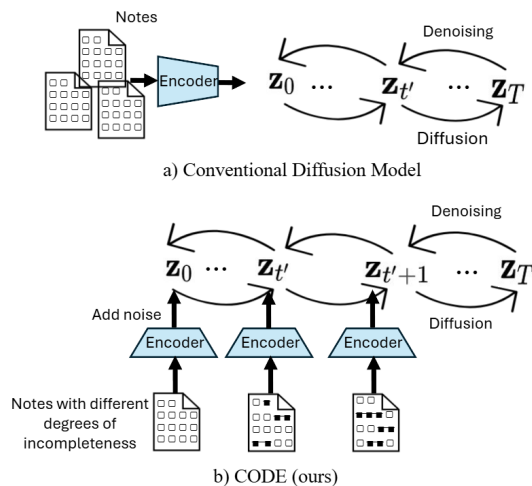


Figure 1: Conventional diffusion model and our proposed CODE model. a) A diffusion model learns to generate encoded complete clinical notes. b) CODE learns to align the diffusion process with the encoded clinical notes with incompleteness so that the missing semantics can be recovered via the denoising process.

detect (Wu et al., 2023), which can be triggered by under-reporting or non-existence of events.

Existing approaches to handling missing structured data (Zhang et al., 2023a; Utama et al., 2024) fail to address missingness in unstructured data such as nursing notes. Also, supervised text-completion methods (Lee et al., 2025; Horvitz et al., 2024) are not applicable to our setting as the ground truth “complete” notes are required. Large-language models (LLMs), though powerful, show serious hallucination in our experiments. Soni and Demner-Fushman (2024) demonstrates the possibility of generating nursing progress notes based on time series and previous reports, leaving the incompleteness issue of clinical notes untouched. There also exist methods to correct errors or inconsistent expressions (Ben Abacha et al., 2025; Asiimwe et al., 2024; Murata et al., 2024), but not explicitly for incompleteness.

To bridge this gap, we propose Completeness-

oriented Diffusion Embeddings (CODE)¹, a novel diffusion-based (Ho et al., 2020) model which generates note embeddings following the distribution of the embeddings of complete notes. The key idea behind CODE is to *align* the embeddings of incomplete notes with the intermediate states of the diffusion process (Fig. 1) via contrastive learning, so the semantics of a potentially complete note are recovered by denoising. To fuse time series with notes, CODE takes time series embeddings as a diffusion condition.

We conduct experiments using data extracted from a real-world EHR dataset (MIMIC-III) for three different prediction tasks including in-hospital mortality, length-of-stay, and septic-shock classification. Experimental results demonstrate that CODE with the ability to recover the missing semantics of clinical notes can achieve higher prediction accuracy across multiple clinical benchmarks. The main contributions of this paper include:

- We propose a novel diffusion-based model (CODE) to generate the embeddings of complete semantics from the embeddings of less complete semantics.
- We extend CODE by taking time-series embeddings to be the diffusion condition and design contrastive learning to align diffusion process with note incompleteness.
- We demonstrate the viability of recovering the embeddings of complete semantics, which leads to performance improvement of three downstream tasks.

2 Related Work

Fusion of Clinical Notes and Time Series

Various methods have been proposed to learn fused representations from time series and clinical notes, where the underlying fusion challenges include temporally misaligned records and insufficient contextual information. In particular, Zhang et al. (2023b) tackles the irregularity of timestamps by querying temporally irregular key-value pairs using regular time intervals. Han et al. (2024) adopts a mixture-of-experts (MoE) approach to achieve data fusion by routing each modality to designated experts that specialize in the modality. For better fusion, Zhu et al. (2024) augments contextual

information by retrieving knowledge based on multimodal data using the RAG approach, which leverages LLMs. Yun et al. (2025) explores inter-modal and intra-modal contexts to supplement missing modalities for the fusion. These methods do not focus explicitly on the incompleteness issue of notes.

Diffusion Models for Text Generation

Diffusion models (Ho et al., 2020) consist of a forward diffusion process of adding noise and a reverse denoising process, which are trained based on a dataset to generate new data that will follow the training data distribution. For diffusion language models, Lovelace et al. (2023) validates that continuous diffusion models can be learned in the latent space of language autoencoders. Shi et al. (2024) unlocks the potential of masked diffusion models for text generation by enabling their training with state-dependent masking schedules. They have also been extended for controlled text generation (Mirbeygi and Beigy, 2025) and text editing (Lee et al., 2025).

In contrast, our proposed method adds noise to the embeddings adaptively based on an estimated degree of *incompleteness*, and tries to train the diffusion model to generate outputs with the missing semantics recovered.

3 Method

3.1 Notations and Prediction Tasks

Assume an EHR dataset of K patients where the record of each patient i contains: a time series of N_i clinical features $\mathbf{X}_i^{ts} := \{(\mathbf{x}_1^{ts}, t_1^{ts}), (\mathbf{x}_2^{ts}, t_2^{ts}), \dots, (\mathbf{x}_{N_i}^{ts}, t_{N_i}^{ts})\}$, a set of M_i time-stamped nursing notes $\mathbf{X}_i^{txt} := \{(\mathbf{x}_1^{txt}, t_1^{txt}), (\mathbf{x}_2^{txt}, t_2^{txt}), \dots, (\mathbf{x}_{M_i}^{txt}, t_{M_i}^{txt})\}$, and a binary class label $y_i \in \{0, 1\}$.

Our goal is to train a model which can extract the information from $(\mathbf{X}_i^{ts}, \mathbf{X}_i^{txt})$ and fuse them to predict y_i . We consider that each nursing note may have an unknown degree of incompleteness, which will cause undesirable bias in model training. We focus on three prediction tasks: in-hospital mortality (IHM), length of stay (LOS), and septic-shock classification (SSC). Predicting IHM is important for treatment plan adjustment if the prediction indicates that a patient has a high risk of death. LOS is to predict whether a patient will be discharged in upcoming days, which can indicate the effectiveness of treatment. SSC is to distinguish septic shock from sepsis without shock, which is

¹<https://github.com/FeilongWu/CODE>

error-prone as they share many symptoms in common (Singer et al., 2016; Schorr et al., 2014).

3.2 Modeling Incomplete Notes Using Diffusion Models

We argue that when an incomplete note with information loss becomes less dependent on the corresponding complete version. The process of losing information is analogous to the diffusion process in which an intermediate state gradually loses dependence on its initial state as the diffusion schedule t increases. In this sense, losing textual information amounts to increasing the diffusion schedule t . This inspires us to model the degree of incompleteness as a diffusion process so that the data of higher degree of completeness can be generated through denoising. As we focus on the semantics of notes for the downstream prediction tasks, we encode the diffusion state using the [CLS] embedding of text, which is extracted from a pretrained encoder (e.g., GatorTron (Yang et al., 2022)).

Our idea of data generation may seem similar to the controlled diffusion generation (Lee et al., 2025; Meng et al., 2022). For example, SDEdit (Meng et al., 2022) generates photo-realistic images from hand-drawings. The key is that the distribution of images can match that of hand-drawings when the same amount of noise is added to both; therefore, images can be generated from noisy hand-drawings through denoising. Yet, direct application to our case is not possible. First, the amount of noise to be added should vary according to the degree of incompleteness rather than being pre-defined and fixed as in (Lee et al., 2025; Meng et al., 2022). Aligning an unknown degree of incompleteness to the amount of noise or diffusion schedule t is non-trivial. In addition, lacking the embeddings of complete notes as the generation targets results in additional challenges.

3.3 Aligning Incomplete Notes and Diffusion Steps via Contrastive Loss

Given a nursing note $\mathbf{x}^{txt}$, we use a pretrained encoder to obtain the [CLS] embedding as its representation, denoted as $\mathbf{z}^{BT}(\mathbf{x}^{txt})$ (or simply $\mathbf{z}^{BT}$). We use GatorTron (Yang et al., 2022) since its token limit is 2,048, which is longer than all notes in our dataset, to ensure no information loss due to the token limit.

To match it with an incompleteness level represented by a diffusion step, $\mathbf{z}^{BT}$ is projected to an

intermediate state $\mathbf{z}_{t'}$ as:

$$p(\mathbf{z}_{t'}|t', \mathbf{z}^{BT}) := \mathcal{N}(\mathbf{z}_{t'}; \sqrt{\bar{\alpha}_{t'}}\mathbf{z}^{BT}, (1 - \bar{\alpha}_{t'})\mathbf{I}),$$

where $\alpha_0 = 1$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, $\{\alpha_i\}$ are fixed hyperparameters, and t' is sampled from a learnable integer distribution $t' \sim p_\psi(t'|\mathbf{z}^{BT})$ with support $t' \in \{0, 1, \dots, T\}$. We use an MLP parameterized by ψ to implement $p_\psi(t'|\mathbf{z}^{BT})$ for aligning the note embedding $\mathbf{z}^{BT}$ with the corresponding diffusion step t' .

To learn $p_\psi(t'|\mathbf{z}^{BT}(\mathbf{x}))$, we propose a contrastive loss which encourages it to give a higher value of t' if the degree of incompleteness of $\mathbf{x}$ is higher. Intuitively, we expect that $\mathbf{z}_{t'}$ will differ more from the inferred $\mathbf{z}_0$ when more sentences are deleted from the note $\mathbf{x}^{txt}$. We start with a note $\mathbf{x}^{txt}$ and randomly delete some sentences from it to create another pseudo note $\mathbf{x}^{txt_{inc}}$ which should have a higher degree of incompleteness. Compared with $\mathbf{x}^{txt}$, the representation of $\mathbf{x}^{txt_{inc}}$ should take more denoising steps to reach $t' = 0$ which should correspond to the embedding of the complete note. In other words, we expect $\mathbb{E}[t'|\mathbf{z}^{BT}(\mathbf{x}^{txt})] < \mathbb{E}[t'|\mathbf{z}^{BT}(\mathbf{x}^{txt_{inc}})]$. Thus, we define the contrastive loss as follows:

$$\mathcal{L}_{contrast} = \frac{1}{|\mathbf{S}|^2} \sum_{a \in \mathbf{S}} \sum_{b \in \mathbf{S}} w(a, b) (t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}}))), \quad (1)$$

where

$$t'_*(\mathbf{z}^{BT}(\mathbf{x})) = \frac{1}{T} \mathbb{E}[t'|\mathbf{z}^{BT}(\mathbf{x})] = \frac{1}{T} \sum_{t'=0}^T t' \cdot p_\psi(t'|\mathbf{z}^{BT}(\mathbf{x}))$$

is the expected degree of incompleteness, which is normalized between 0 and 1, and $\mathbf{S}$ denotes the index set of all notes that $\mathbf{S} := \{i_j : i \in \{1, 2, \dots, K\}, j \in \{1, 2, \dots, M_i\}\}$.

The proposed weighting function is as:

$$w(a, b) = \begin{cases} 1 & \text{if } a = b \\ \frac{h(\mathbf{x}_b^{txt_{inc}}, \mathbf{x}_a^{txt})}{\sum_{b \in \mathbf{S}} h(\mathbf{x}_b^{txt_{inc}}, \mathbf{x}_a^{txt})} & \text{else,} \end{cases} \quad (2)$$

$$h(\mathbf{x}_b^{txt_{inc}}, \mathbf{x}_a^{txt}) = \lambda_1 \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) + (1 - \lambda_1) \log (1 - \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt}))) \quad (3)$$

with $0 < \lambda_1 < 1$. Note that for $w(a, b)$, when $a = b$ (i.e., the same note), the incompleteness level of the incomplete version should always be larger than its original version's, and thus we set $w(a, b) = 1$ (a.k.a. positive pairs in contrastive

learning) (Chen et al., 2020). Considering only these positive pairs is insufficient as it will force t'_* either close to 0 or 1 according to Eq.(1) which is undesirable. We also consider pairs of different notes (i.e., $a \neq b$) and use soft weighting (Park et al., 2024) for $w(a, b)$ to implicitly align the value of t'_* across the notes, which automatically adjusts its sign when the t'_* of two notes in a pair differ. We find that adjusting λ_1 can control the magnitude of the learned t'_* , where higher λ_1 may favor a large gap in the completeness level between the given and incomplete notes.

3.4 Diffusion for Complete Semantics

Once $\mathbf{z}^{BT}$ is mapped to $\mathbf{z}_{t'}$ via the contrastive module, the diffusion process splits into two ranges that require different training signals: a supervised range $[t', T]$, where adding Gaussian noise to $\mathbf{z}_{t'}$ provides exact denoising targets, and an unsupervised range $[0, t']$, where the target $\mathbf{z}_0$, the embedding of an unobserved complete note, is unknown.

For the supervised range, we adopt the standard DDPM (Ho et al., 2020) loss:

$$\mathcal{L}_{t':T} = \mathbb{E}_{\mathbf{z}_{t'}, \epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_{t'+1:t}} \mathbf{z}_{t'} + \sqrt{1 - \bar{\alpha}_{t'+1:t}} \epsilon, t \right) \right\|^2 \right], \quad (4)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \text{Uniform}(t', T)$; the variational-bound derivation is in Appendix A.

For the unsupervised range, since $\mathbf{z}_0$ is unobserved, we instead regularize the predicted noise to be consistent with the prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathcal{L}_{0:t'} = \mathbb{E}_{\mathbf{z}_{1:t'}, \epsilon} \left[\frac{1}{t'} \sum_{t=1}^{t'} \left\| \epsilon_\theta(\mathbf{z}_t, t) \right\|^2 + \left\| \text{cov}(\epsilon_\theta(\mathbf{z}_t, t)) - \mathbf{I} \right\|^2 \right], \quad (5)$$

where $\text{cov}(\cdot)$ denotes the empirical covariance over the minibatch. This regularization discourages arbitrary noise predictions in the unsupervised range, which would otherwise disturb the recovery of $\mathbf{z}_0$ enforced by $\mathcal{L}_{\text{complete}}$ (Sec. 3.5). The full denoising objective is:

$$\mathcal{L}_{\text{denoise}} := \lambda_0 \mathcal{L}_{t':T} + (1 - \lambda_0) \mathcal{L}_{0:t'}, \quad (6)$$

where λ_0 is a weighting factor to balance between true noise and regularization.

During inference, we compute the expectation of the latent variable $\mathbf{z}_0$ which corresponds to the

pseudo complete note embedding:

$$\begin{aligned} & \mathbb{E}_{\epsilon, t'} [p_\theta(\mathbf{z}_0 | \mathbf{z}^{BT})] \\ &= \mathbb{E}_\epsilon \left[\sum_{t'} p_\psi(t' | \mathbf{z}^{BT}) p(\mathbf{z}_{t'} | t', \mathbf{z}^{BT}) \prod_{t=1}^{t'} p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t) \right]. \end{aligned}$$

We use the results from (Ho et al., 2020) that

$$p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_t^2 \mathbf{I}),$$

where

$$\begin{aligned} \mu_\theta(\mathbf{z}_t, t) &= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{z}_t, t) \right), \\ \sigma_t^2 &= \frac{(1 - \bar{\alpha}_t)(1 - \alpha_t)}{1 - \bar{\alpha}_t}. \end{aligned}$$

3.5 Anchoring $\mathbf{z}_0$ Generation

While the complete notes are unknown, we guide the estimation of $\mathbf{z}_0$ close to the complete note embedding by imposing desirable properties. Specifically, we design a loss term to fulfill the following two criteria related to $\mathbf{z}_0$:

- (i) **Invariance to incompleteness.** The reverse process should recover the same expected $\mathbf{z}_0$ from a note $\mathbf{x}^{txt}$ and its corrupted version $\mathbf{x}^{txt_{inc}}$;
- (ii) **Anchoring on confidently-complete inputs.** When the contrastive module judges $\mathbf{x}^{txt}$ to be already complete (i.e., $p_\psi(t' = 0 | \mathbf{z}^{BT})$ greater than a threshold), the recovery from its corrupted counterpart should remain close to $\mathbf{z}^{BT}$ itself.

Denoting $\mathbf{z}^{BT}$ and $\mathbf{z}_{inc}^{BT}$ for the encoder outputs of $\mathbf{x}^{txt}$ and $\mathbf{x}^{txt_{inc}}$, and $\mathbf{z}_{t'}, \mathbf{z}_{t', inc}$ for their forward-diffusion samples, we combine the two principles into a single loss:

$$\begin{aligned} \mathcal{L}_{\text{complete}} &= \\ & \left\| \mathbb{E}_{p_\psi(t' | \mathbf{z}^{BT}), \epsilon} [\mathbf{z}_0 | \mathbf{z}_{t'}] - \mathbb{E}_{p_\psi(t' | \mathbf{z}_{inc}^{BT}), \epsilon} [\mathbf{z}_0 | \mathbf{z}_{t', inc}] \right\|^2 \\ & + g(\mathbf{x}^{txt}, \lambda_2) \left\| \mathbf{z}^{BT} - \mathbb{E}_{p_\psi(t' | \mathbf{z}_{inc}^{BT}), \epsilon} [\mathbf{z}_0 | \mathbf{z}_{t', inc}] \right\|^2, \end{aligned} \quad (7)$$

where $g(\mathbf{x}^{txt}, \lambda_2) = \mathbb{1} [p_\psi(t' = 0 | \mathbf{z}^{BT}(\mathbf{x}^{txt})) \geq \lambda_2]$ is an indicator function that activates the anchoring term only when the input is judged confidently complete, with threshold $\lambda_2 \in [0, 1]$. The first term enforces principle (i); the second, gated by $g(\cdot)$, enforces principle (ii) and provides the only direct supervision on $\mathbf{z}_0$, addressing the absence of ground-truth complete notes. At inference time, the same threshold acts as a gate: embeddings with $p_\psi(t' = 0 | \mathbf{z}^{BT}) \geq \lambda_2$ are treated as already complete and bypass the diffusion model, while the remaining incomplete embeddings are passed through to produce $\mathbf{z}_0$.

3.6 Conditioning Diffusion on Time-series for Prediction

To extend our diffusion model to incorporate time series and facilitate multimodal fusion, we parameterize the model as $\epsilon_\theta(\mathbf{z}_{t'}, \mathbf{e}^{ts}, t')$. $\mathbf{e}^{ts}$ is the embedding of the time series used as the diffusion model's condition obtained by mTAND (Shukla and Marlin, 2021), given as:

$$\mathbf{e}^{ts} = \text{mTAND} \left(\text{Q}(t^{txt}), \text{K}([t_1^{ts}, t_2^{ts}, \dots, t_M^{ts}]), \text{V}([\mathbf{x}_1^{ts}, \mathbf{x}_2^{ts}, \dots, \mathbf{x}_M^{ts}]) \right), \quad (8)$$

where the query Q converts the timestamps of the nursing notes corresponding to $\mathbf{z}_{t'}$ into embeddings. Similarly, both keys K and values V are embeddings using both time-series timestamps and time-series, respectively.

For the prediction head, we adopt similar modules of the SOTA model MoE-Retriever (Yun et al., 2025):

$$\hat{y} = \text{Classifier}(f^{ts} \otimes f^{txt}),$$

where $\otimes$ denotes concatenation and

$$\begin{aligned} f^{ts} &= F^{ts} \left((\mathbf{x}_1^{ts}, t_1^{ts}), \dots, (\mathbf{x}_N^{ts}, t_N^{ts}) \right) \\ f^{txt} &= F^{txt} \left((\mathbb{E}_{p(t'|\mathbf{z}_{t'_1}), \epsilon}[\mathbf{z}_{01}|\mathbf{z}_{t'_1}], t_1^{txt}), \dots, (\mathbb{E}_{p(t'|\mathbf{z}_{t'_M}), \epsilon}[\mathbf{z}_{0M}|\mathbf{z}_{t'_M}], t_M^{txt}) \right). \end{aligned}$$

$F^{ts}(\cdot)$ is a network that takes a time series as input and projects each time series element $\mathbf{x}^{ts}$ and its timestamps t^{ts} onto their corresponding embedding spaces. The two embeddings are then added together with a self-attention layer followed to give a time-series element feature vector. The element feature vectors over the time series are averaged to give f^{ts} . $F^{txt}(\cdot)$ is another network that projects the timestamps of a note $\mathbf{x}^{txt}$ to a time embedding and adds it to the expected $\{\mathbf{z}_0\}$ from the diffusion model, and then again passes it a self-attention layer to give the average feature vector f^{txt} over notes.

To train the classifier, we adopt a cross entropy loss between the ground truth label y and the predicted class label $\hat{y}$:

$$\mathcal{L}_{cls} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})). \quad (9)$$

The overall loss function for training the proposed CODE model becomes:

$$\mathcal{L}_{CODE} = \mathcal{L}_{denoise} + \mathcal{L}_{contrast} + \mathcal{L}_{complete} + \mathcal{L}_{cls} \quad (10)$$

by substituting the denoiser $\epsilon_\theta(\mathbf{z}_t, t)$ with the extended one $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}^{ts}, t)$ for multimodal fusion.

While the model can be trained end-to-end, we find it useful to pretrain the network for $p_\psi(t'|\mathbf{z}^{BT}(\mathbf{x}))$ by first minimizing only $\mathcal{L}_{contrast}$. Then, we freeze the parameters ψ and train the model by minimizing the overall loss function. Fig. 2 depicts our model architecture, with the details of the denoiser provided in Appendix E.1.

4 Experiments

4.1 Data and Evaluation Tasks

We evaluate CODE on three ICU patient-outcome prediction tasks derived from MIMIC-III (Johnson et al., 2016): in-hospital mortality (IHM), length of stay (LOS), and septic-shock classification (SSC). For each patient, we extract a 37-dimensional time series of vital signs and laboratory measurements together with the nursing notes recorded during the ICU stay. IHM and LOS use a 48-hour observation window, with the targets being death before discharge and discharge within the next 48 hours respectively. SSC uses a 72-hour window, with the target being the onset of septic shock. After excluding patients who die within the window or have excessive feature missingness, the cohorts contain 10,973 (IHM; 11.2% positive), 11,472 (LOS; 39.4% positive), and 5,966 (SSC; 10.3% positive) patients. More details can be found in Appendix D.1. As ground-truth complete notes are unavailable, we construct training pairs by randomly deleting 1–12 sentences from each note to obtain its incomplete counterpart $\mathbf{x}^{txt_{inc}}$ used in the contrastive objective (Sec. 3.3); the same protocol is also used at test time to probe robustness against varying levels of incompleteness.

We compare CODE against seven recent multimodal clinical models: MT (Lyu et al., 2023), UTDE (Zhang et al., 2023b), MCP (King et al., 2023), FuseMoE (Han et al., 2024), MoE-Retriever (Yun et al., 2025), AutoFM (Cui et al., 2024), and CTPD (Wang et al., 2025). More details of the baselines can be found in Appendix D.1. GatorTron (Yang et al., 2022) is used as the text encoder. We report F1, AUPRC, and AUROC averaged over five runs with random 6:2:2 train/validation/test splits. Hyperparameter search spaces are summarized in Appendix E.2.

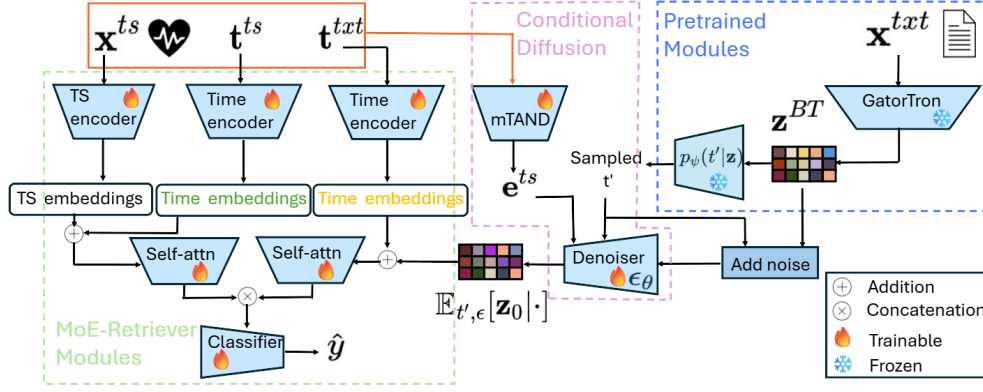


Figure 2: The model architecture of CODE.

Table 1: Performance comparison on three prediction tasks. Metrics are reported as Mean_{Std}.

		MT	UTDE	FuseMoE	MCP	MoE-Retriever	AUTOFM	CTPD	CODE (ours)
IHM	F1	0.386 _{0.052}	0.520 _{0.033}	0.516 _{0.026}	0.322 _{0.033}	<u>0.528_{0.024}</u>	0.517 _{0.031}	0.505 _{0.014}	0.549_{0.017}
	AUPRC	0.410 _{0.030}	0.584 _{0.024}	0.555 _{0.017}	0.370 _{0.019}	<u>0.590_{0.018}</u>	0.573 _{0.021}	0.559 _{0.016}	0.597_{0.021}
	AUROC	0.843 _{0.011}	0.904 _{0.007}	0.897 _{0.006}	0.795 _{0.021}	<u>0.906_{0.005}</u>	0.900 _{0.007}	0.905 _{0.006}	0.912_{0.005}
LOS	F1	0.663 _{0.052}	<u>0.745_{0.013}</u>	0.744 _{0.022}	0.632 _{0.033}	0.739 _{0.010}	0.736 _{0.017}	0.744 _{0.013}	0.757_{0.012}
	AUPRC	0.754 _{0.023}	<u>0.792_{0.015}</u>	0.785 _{0.012}	0.649 _{0.027}	<u>0.803_{0.014}</u>	0.779 _{0.016}	0.791 _{0.016}	0.818_{0.011}
	AUROC	0.828 _{0.007}	0.869 _{0.007}	0.865 _{0.008}	0.740 _{0.028}	<u>0.873_{0.006}</u>	0.862 _{0.006}	0.869 _{0.005}	0.876_{0.006}
SSC	F1	0.391 _{0.030}	0.404 _{0.059}	0.358 _{0.025}	0.232 _{0.031}	0.409 _{0.062}	0.310 _{0.034}	0.402 _{0.025}	0.426_{0.058}
	AUPRC	0.385 _{0.021}	0.425 _{0.036}	0.380 _{0.043}	0.207 _{0.009}	0.429 _{0.045}	0.390 _{0.026}	<u>0.433_{0.047}</u>	0.453_{0.037}
	AUROC	0.806 _{0.003}	0.840 _{0.019}	<u>0.842_{0.028}</u>	0.603 _{0.022}	0.840 _{0.023}	0.822 _{0.015}	<u>0.842_{0.020}</u>	0.860_{0.017}

4.2 Results

Table 1 compares CODE with the seven baselines on the three tasks. CODE consistently achieves the best F1, AUPRC, and AUROC on all tasks. Specifically, for IHM, CODE outperforms the strongest baseline MoE-Retriever by 4.0% based on F1 score. Even greater relative gains are observed on SSC, where CODE achieves AUPRC from 0.433 to 0.453 (4.6%) over the strongest baseline CTPD. This is consistent with the intuition that recovering missing semantics is more effective when class separation relies on subtle cues in nursing notes, as in distinguishing septic shock from sepsis. The same observation can be made when GatorTron is replaced by ClinicalBERT (Fig. 11 in Appendix G.1), indicating that the gains stem from the diffusion-based completion mechanism rather than the choice of encoder. Additional evidence of the gains can be seen by comparing CODE with MoE-Retriever as they share the same model architecture except that the [CLS] embeddings are modified by the diffusion module of CODE. In Appendix G.5, we explore some simple baselines as choices of model architecture, which suggests that diffusion model is effective for our problem.

We further investigate the robustness of CODE

and baseline models by deleting 1–11 sentences from each note. Figure 3 shows that the accuracy baselines degrades steadily on all three tasks as more sentences are removed, whereas CODE remains comparatively flat. This shows the effectiveness of our method: by aligning incomplete-note embeddings with intermediate diffusion states and denoising toward complete-note semantics, the downstream predictor receives a more invariant input distribution across incompleteness levels.

4.3 Ablation Studies of Loss Terms

To further understand the contribution of each loss term in Eq. 10, we conduct an ablation study over the three diffusion-related losses $\mathcal{L}_{denoise}$, $\mathcal{L}_{contrast}$, and $\mathcal{L}_{complete}$, keeping $\mathcal{L}_{cls}$ in all configurations. Table 2 reports the IHM results; the same trends hold on LOS and SSC (Appendix G.2). Among the three losses used in isolation, $\mathcal{L}_{complete}$ contributes the most: on IHM, it raises AUROC from 0.805 to 0.852. $\mathcal{L}_{contrast}$ alone yields only a marginal gain over the no-loss baseline, but is highly complementary to $\mathcal{L}_{complete}$: combining the two reaches AUROC 0.877, because the incompleteness ordering induced by the estimated t' provides valid generation targets for $\mathcal{L}_{complete}$.

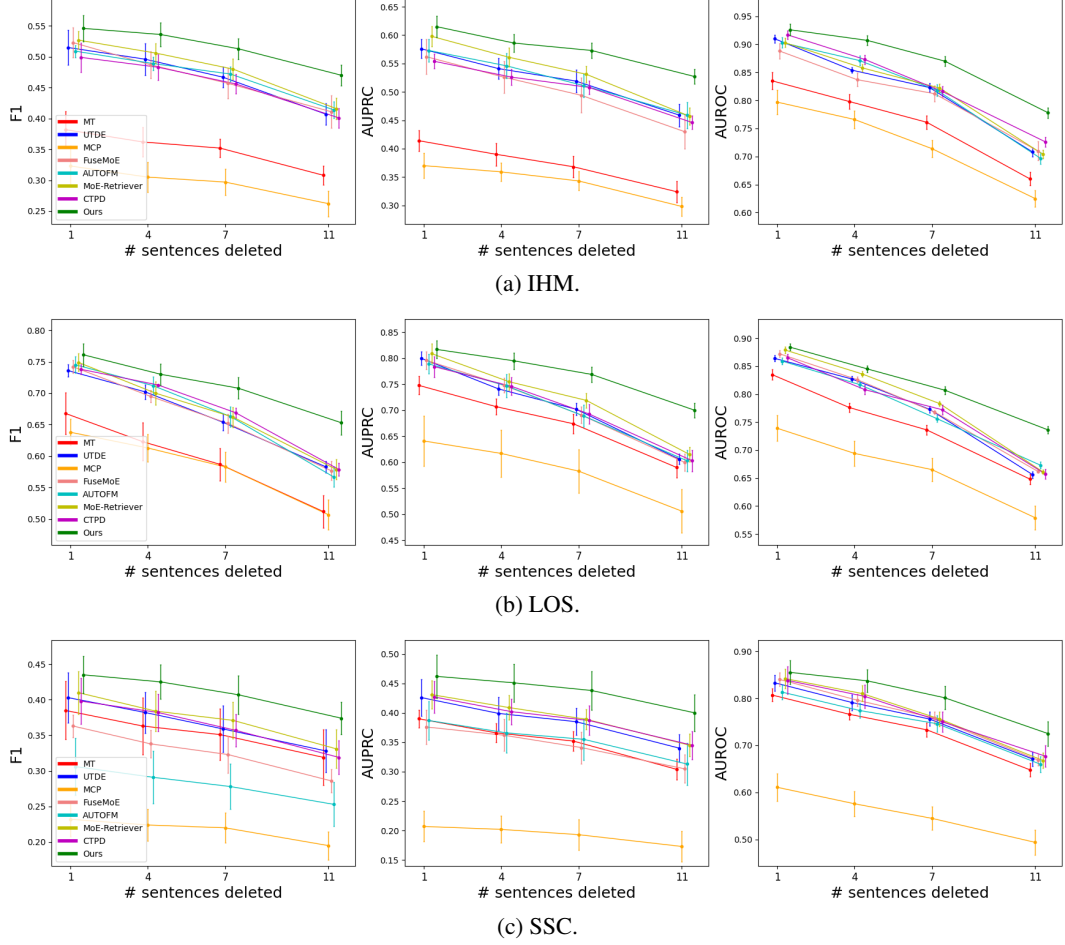


Figure 3: Performance comparison on in-hospital mortality (a), length of stay (b), and septic shock classification (c) with different numbers of sentences deleted from nursing notes.

Adding $\mathcal{L}_{denoise}$ on top yields the full objective and the best AUROC of 0.912; removing the regularization term $\mathcal{L}_{0:t'}$ inside $\mathcal{L}_{denoise}$ drops AUROC to 0.885, confirming that constraining the predicted noise to match $\mathcal{N}(0, \mathbf{I})$ over $[0, t']$ is necessary to prevent arbitrary noise predictions in this unsupervised range.

We further compare conditional and unconditional variants of the diffusion model (Table 3). Removing the time-series condition e^{ts} drops IHM AUROC from 0.912 to 0.881, with consistent degradations on LOS (0.876→0.868) and SSC (0.860→0.825), confirming that the time-series condition is the source of the multimodal-fusion gain.

4.4 Case Study

In addition to conducting downstream evaluation to demonstrate the predictive power of the recovered semantics, we further perform a case study on whether the recovered semantics correspond

to clinically significant information. We select five patients from the IHM dataset, for whom the strongest baselines including MoE-Retriever, UTDE, CTPD all make a correct prediction. We gradually delete clinically-important sentences suggested by an LLM for their nursing notes until these baselines make an incorrect prediction while the proposed model’s prediction remains correct for the corrupted notes. We invite an experienced ICU clinician to determine if each removed sentence contains i) “important facts”, ii) “interpretations” of some facts, or iii) information with “uncertain importance”. Out of the 28 deleted sentences in the 5 cases, there are 11 (39%) important facts and 8 (29%) interpretations, where one of them is a radiological diagnosis. The remaining 9 (32%) sentences are of uncertain importance. This provides further evidence (though not direct) that CODE is capable of recovering missing semantics which are important for the downstream prediction task. One of the five cases can be found in Appendix F.

Table 2: Results of ablation study on IHM prediction task.

$+\mathcal{L}_{denoise}$	$+\mathcal{L}_{contrast}$	$+\mathcal{L}_{complete}$	F1	AUPRC	AUROC
✗	✗	✗	0.460 _{0.015}	0.498 _{0.013}	0.805 _{0.009}
✓	✗	✗	0.463 _{0.017}	0.526 _{0.022}	0.834 _{0.008}
✗	✓	✗	0.461 _{0.028}	0.502 _{0.010}	0.808 _{0.010}
✗	✗	✓	0.483 _{0.013}	0.542 _{0.014}	0.852 _{0.010}
-----			-----		
✓	✓	✗	0.506 _{0.027}	0.565 _{0.013}	0.869 _{0.011}
✓	✗	✓	0.501 _{0.024}	0.555 _{0.019}	0.863 _{0.010}
✗	✓	✓	0.519 _{0.020}	0.569 _{0.016}	0.877 _{0.009}
-----			-----		
✓(no $\mathcal{L}_{0:t'}$)	✓	✓	0.532 _{0.021}	0.578 _{0.013}	0.885 _{0.008}
✓	✓	✓	0.549 _{0.017}	0.597 _{0.021}	0.912 _{0.005}

Table 3: Ablation study of the diffusion with the encoder as GatorTron. Metrics are reported as Mean_{Std}.

		F1	AUPRC	AUROC
IHM	Unconditional Diffusion	0.513 _{0.026}	0.563 _{0.017}	0.881 _{0.010}
	Conditional Diffusion	0.549 _{0.017}	0.597 _{0.021}	0.912 _{0.005}
LOS	Unconditional Diffusion	0.719 _{0.012}	0.778 _{0.015}	0.868 _{0.009}
	Conditional Diffusion	0.757 _{0.010}	0.818 _{0.011}	0.876 _{0.006}
SSC	Unconditional Diffusion	0.394 _{0.052}	0.424 _{0.028}	0.825 _{0.017}
	Conditional Diffusion	0.426 _{0.058}	0.453 _{0.037}	0.860 _{0.017}

4.5 Effectiveness on Recovering Complete Note Embedding

We further attempt to evaluate explicitly how well it can recover the embedding of a complete note from an incomplete nursing note. Since we do not have ground truth notes in the real-world datasets, we create a synthetic one consisting of 10K nursing progress notes using DeepSeek-R1 (DeepSeek-AI, 2025) to conduct this experiment. We assume these synthetic notes are the complete ground truth, which is ensured by a carefully-designed prompt with instructions to include specific sections in the notes (see Appendix D.2 for details on synthetic data generation). To introduce incompleteness to a note, we randomly delete one to seven sentences. We create three different datasets with different proportions (40%, 60%, and 80%) of incomplete notes. We use the embeddings obtained by applying only ClinicalBERT (called “Naive”) as the baseline and compare the embeddings obtained using the CODE model. We use cosine distance to measure how close they are to the embeddings obtained by applying ClinicalBERT to the ground truth notes. According to Figure 4, CODE can recover the embeddings of different levels of semantic completeness closer to the ground truth ones when compared with the “Naive” baseline, which is in fact commonly used in predictive analytics on clinical notes.

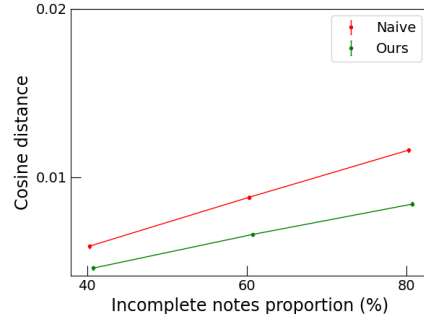


Figure 4: Evaluating CODE for recovering the [CLS] embeddings w.r.t ground truth embeddings.

4.6 How about Recovering Complete Notes Using LLMs?

The proposed model addresses the incompleteness issue via latent representations. To see if one can simply use LLMs to address the same problem, we tested ChatGPT-5 (OpenAI, 2026), Gemini-2.5-Pro (Google, 2024), DeepSeek-R1 (DeepSeek-AI, 2025), and Qwen-3-Max (Yang et al., 2025) with a carefully designed prompt (see Appendix E.3). We find that hallucination is a serious issue for this completion task. Examples are in Appendix G.6. Specifically, LLMs are unaware of the missing information or generate responses deviating from ground truth. This makes the semantics of the LLM-generated notes even far away from the ground truth (see Appendix G.3).

Method	Inference Latency (sec.)
CTPD	1.08
AutoFM	0.25
MoE-Retriever	0.15
FuseMoE	1.14
CODE (ours)	2.45

Table 4: Inference latency per patient from IHM.

	$T = 20$	$T = 10$
F1	0.547 _{0.018}	0.545 _{0.014}
AUPRC	0.596 _{0.020}	0.594 _{0.018}
AUROC	0.910 _{0.005}	0.909 _{0.004}
Latency (sec.)	1.34	0.81

Table 5: The performance of CODE on IHM with accelerated inference latency using DDIM (Song et al., 2021).

4.7 Can the Estimated t' Represent Completeness?

The estimated value of $p_\psi(t' = 0|\mathbf{z})$ is used to select embeddings of high levels of completeness as generation targets. We rank them in a descending order and show their corresponding notes from MIMIC-III dataset. We can notice that the degree of completeness drops as $p_\psi(t' = 0|\mathbf{z})$ drops (see Appendix G.6). Furthermore, the average value of t'_* for the embeddings of our model generation, given notes, and incomplete notes increases from left to right, aligning with our expectation that a higher value of t'_* indicates a higher level of incompleteness (see Appendix G.3).

4.8 Inference Latency

We use one NVIDIA V100-32G to test the inference latency of the proposed method ($T = 40$) per patient with 48-hr observation data. Table 4 shows that the inference latency of our method is higher than other baselines, primary due to the use of diffusion model. To accelerate the inference, we can use less samples during the denoising using DDIM (Song et al., 2021), but the performance will be slightly undermined (Table 5).

5 Conclusion

We propose a novel diffusion-based model for fusing clinical notes and time series data for healthcare analytics with superior performance. We validate its capability to recover embeddings of high semantic completeness from those of lower semantic

completeness, which provides a solution to the issue of learning representations from incomplete clinical notes.

Limitations

CODE is designed with the expectation that $p_\psi(t' = 0|\mathbf{z}^{BT})$ is close to 1 if a note is nearly complete. Empirically, it is often hard to achieve, which motivates our design of using a threshold function in Eq. 7. Future extension of this work can focus on designing better contrastive loss or using more fine-grained text representation beyond [CLS] embedding.

Furthermore, this work assumes that the missingness is due to random sentence-omission. Since real missing patterns can be different from random omission and may depend on clinical significance, we further explored random omission only on clinically significant or less significant information. The results are suboptimal to the proposed unrestricted random omission (Appendix G.4). Other missing patterns such as word omission or brief narration could be investigated in future work.

Acknowledgments

The authors thank Dr. Kenny Chan for providing advice for the interpretations of the clinical notes. This work is partially supported by the National Natural Science Foundation of China (62302413) and the Health and Medical Research Fund (23220312).

References

- Arnold Caleb Asiimwe, Didac Suris Coll-Vinent, Pranav Rajpurkar, and Carl Vondrick. 2024. MedAutoCorrect: Image-conditioned autocorrection in medical reporting. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252.
- Asma Ben Abacha, Wen-wai Yim, Yajuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550, Vienna, Austria.
- Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. 2013. Generalized denoising auto-encoders as generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'13*, page 899–907, Red Hook, NY, USA. Curran Associates Inc.
- Liz Charalambous and Sarah Goldberg. 2016. [Gaps, mishaps and overlaps. Nursing documentation: How](#)

- does it affect care? *Journal of Research in Nursing*, 21.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*.
- Arman Cohan, Luca Soldaini, Nazli Goharian, Allan Fong, Ross Filice, and Raj Ratwani. 2016. [Identifying significance of discrepancies in radiology reports](#).
- Suhan Cui, Jiaqi Wang, Yuan Zhong, Han Liu, Ting Wang, and Fenglong Ma. 2024. [Automated fusion of multimodal electronic health records for better medical predictions](#). In *Proceedings of the SIAM International Conference on Data Mining*, volume 2024, pages 361–369.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Google. 2024. [Gemini 2.5 pro](#). Accessed: 2026-02-01.
- Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. 2024. [Fusimoe: Mixture-of-experts transformers for fleximodal fusion](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 67850–67900. Curran Associates, Inc.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denoising diffusion probabilistic models](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, page 18216–18224.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. [Clinicalbert: Modeling clinical notes and predicting hospital readmission](#). Preprint, arXiv:1904.05342.
- Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. [Mimic-iii, a freely accessible critical care database](#). *Scientific Data*, 3:160035.
- Ryan King, Tianbao Yang, and Bobak Mortazavi. 2023. Multimodal pretraining of medical time series and notes. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 1–12. PMLR.
- Che Hyun Lee, Heeseung Kim, Jiheum Yeom, and Sungro Yoon. 2025. [EdiText: Controllable coarse-to-fine text editing with diffusion language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22798–22815, Vienna, Austria.
- Sicen Liu and Hao Chen. 2024. [Knowledge injected multimodal irregular ehrrs model for medical prediction](#). In *Trustworthy Artificial Intelligence for Healthcare: Second International Workshop, TAI4H 2024, Jeju, South Korea, August 4, 2024, Proceedings*, page 25–39, Berlin, Heidelberg. Springer-Verlag.
- Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. 2023. [Latent diffusion for language generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 56998–57025. Curran Associates, Inc.
- Calvin Luo. 2022. [Understanding diffusion models: A unified perspective](#). Preprint, arXiv:2208.11970.
- Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. 2023. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. *AMIA Annual Symposium Proceedings*, 2022:719–728.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proceedings of International Conference on Learning Representations*.
- Mohaddeseh Mirbeygi and Hamid Beigy. 2025. Prompt guided diffusion for controllable text generation. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 78–84, Albuquerque, New Mexico, USA.
- Wataru Murata, Chang Liu, Goshiro Yamamoto, Yukiko Mori, and Tomohiro Kuroda. 2024. [Similarity analysis of clinical notes for revealing writing styles in an electronic medical record system](#). *Studies in health technology and informatics*, 316:1724–1728.
- OpenAI. 2026. Openai api (model: Chatgpt-5). Online; available at <https://platform.openai.com>. Accessed 2026-02-01.
- Minsu Park, Seyeon Choi, Chanyeol Choi, Jun-Seong Kim, and Jy-yong Sohn. 2024. Improving multi-lingual alignment through soft contrastive learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 138–145.
- Prasad Patil. Healthcare dataset. Available at <https://www.kaggle.com/datasets/prasad22/healthcare-dataset>. Collected in 2025.

- Christa A Schorr, Sergio Zanotti, and R Phillip Dellinger. 2014. [Severe sepsis and septic shock: management and performance improvement](#). *Virulence*, 5(1):190—199.
- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. [Simplified and generalized masked diffusion for discrete data](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 103131–103167. Curran Associates, Inc.
- Satya Narayan Shukla and Benjamin Marlin. 2021. [Multi-time attention networks for irregularly sampled time series](#). In *Proceedings of International Conference on Learning Representations*.
- Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. 2016. [The third international consensus definitions for sepsis and septic shock \(sepsis-3\)](#). *JAMA*, 315(8):801–810.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. [Denoising diffusion implicit models](#). In *International Conference on Learning Representations*.
- Sarvesh Soni and Dina Demner-Fushman. 2024. [Toward relieving clinician burden by automatically generating progress notes using interim hospital data](#). *Preprint*, arXiv:2410.12845.
- Agung Bella Putra Utama, Aji Prasetya Wibawa, Anik Nur Handayani, Wahyu Sakti Gunawan Irianto, Aripriharta, and Amat Nyoto. 2024. [Improving time-series forecasting performance using imputation techniques in deep learning](#). In *Proceedings of 2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, pages 232–238.
- Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. 2025. [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, Vienna, Austria. Association for Computational Linguistics.
- Kevin Wu, Dominik Dahlem, Christopher Hane, Eran Halperin, and James Zou. 2023. Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209, pages 229–242.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Shin, Kaleb Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony Costa, Mona Flores, Ying Zhang, Tanja Magoc, Christopher Harle, Gloria Li-pori, Duane Mitchell, William Hogan, Elizabeth Shenkman, Jiang Bian, and Yonghui Wu. 2022. [A large language model for electronic health records](#). *npj Digital Medicine*, 5.
- Sukwon Yun, Jiayi Xin, Inyoung Choi, Jie Peng, Ying Ding, Qi Long, and Tianlong Chen. 2025. [Generate, then retrieve: Addressing missing modalities in multimodal learning via generative AI and moe](#). In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.
- Ke Zhang, Ke Niu, Yuhang Zhou, Wenjuan Tai, and Guoqiang Lu. 2023a. [MedCT-BERT: Multimodal mortality prediction using medical ConvTransformer-BERT model](#). In *Proceedings of 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 700–707.
- Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. 2023b. Improving medical predictions by irregular multimodal electronic health records modeling. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. [Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM ’24*, page 3549–3559, New York, NY, USA. Association for Computing Machinery.

A Derivation

A comprehensive reference for understanding diffusion models is (Luo, 2022). The variational bound of $\log \int p_\theta(\mathbf{z}_{t':T}) d\mathbf{z}_{t'+1:T}$ is similar to vanilla diffusion models (Ho et al., 2020) if $\mathbf{z}_{t'}$ is considered $\mathbf{z}_0$. Therefore,

$$\begin{aligned}
-\log \int p_\theta(\mathbf{z}_{t':T}) d\mathbf{z}_{t'+1:T} &\leq \mathbb{E}_{q(\mathbf{z}_{t'+1:T}|\mathbf{z}_{t'})} \left[D_{KL}(q(\mathbf{z}_T|\mathbf{z}_{t'})||p(\mathbf{z}_T)) + \sum_{t=t'+2}^T (D_{KL}(q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_{t'})) \right. \\
&\quad \left. - \log p_\theta(\mathbf{z}_{t'}|\mathbf{z}_{t'+1})) \right] \\
&\propto \mathbb{E}_{\mathbf{z}_{t'}, \epsilon} \left[\frac{\beta_t}{2\alpha_t(1 - \bar{\alpha}_{t'+1:t-1})} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_{t'+1:t}}\mathbf{z}_{t'} + \sqrt{1 - \bar{\alpha}_{t'+1:t}}\epsilon, t)\|^2 \right] \\
&\propto \mathbb{E}_{\mathbf{z}_{t'}, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_{t'+1:t}}\mathbf{z}_{t'} + \sqrt{1 - \bar{\alpha}_{t'+1:t}}\epsilon, t)\|^2 \right]
\end{aligned} \tag{11}$$

B Contrastive Learning Failure Attempts

We have tested different plausible training objectives for contrast learning, but none of them is comparable to the proposed one in terms of performance. In this subsection, we show some failure cases. We artificially create incomplete notes using the synthetic notes by progressively deleting 2 sentences. Specifically, we are first given a complete note. The second note is created by randomly deleting 2 sentences from the first one. The third note is created by randomly deleting 2 sentences from the second one. In total, three incomplete notes are created for each complete note. For the training data, we randomly delete 0-6 sentences for each note. The test data are the complete notes with their three incomplete notes of progressive deletion. All notes are converted into [CLS] embeddings. The prediction task is to rank the degree of completeness for each complete note paired with its incomplete notes of progressive sentence deletion. For example, the correct order of completeness is [1,2,3,4] for each complete note with its corresponding incomplete notes of progressive deletion. We use the metrics Kendall Rank Correlation Coefficient (τ) and Spearman Rank Coefficient (ρ) to evaluate our prediction of ranking. We determine the ranking based on the estimated t'_* that the smaller the value the higher the completeness. Table 6 shows that the proposed contrastive learning best captures note incompleteness.

Objective 1: This objective features the use of absolute value in the denominator of w , which makes the sign of w only depend on the magnitude of estimated t'_* . If the sign is positive, it indicates the incomplete note should have high degree of incompleteness; otherwise, the sign will be negative and the loss is minimized in an reversed direction.

$$\mathcal{L}_{contrast} = \frac{-1}{|\mathbf{S}|^2} \sum_a \sum_b w(a, b) (\lambda_1 \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) + (1 - \lambda_1) \log (1 - \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt}))),$$

where $0 < \lambda_1 < 1$,

$$t'_*(\mathbf{z}^{BT}(\mathbf{x})) = \frac{1}{T} \sum_{s=0}^T s \cdot p_\psi(t'| \mathbf{z}^{BT}(\mathbf{x}))$$

is the expected degree of incompleteness given $\mathbf{z}^{BT}(\mathbf{x})$ normalized between 0 and 1.

$$w(a, b) = \begin{cases} 1 & a = b \\ \frac{(t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})))}{\sum_b (t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})))} & else. \end{cases}$$

Objective 2: This objective is very similar to the previous one except that exponential function replaces the absolute value function to make the denominator of w positive.

$$\mathcal{L}_{contrast} = \frac{-1}{|\mathbf{S}|^2} \sum_a \sum_b w(a, b) (\lambda_1 \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) + (1 - \lambda_1) \log (1 - \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt}))),$$

where $0 < \lambda_1 < 1$,

$$t'_*(\mathbf{z}^{BT}(\mathbf{x})) = \frac{1}{T} \sum_{s=0}^T s \cdot p_\psi(t'|\mathbf{z}^{BT}(\mathbf{x}))$$

is the expected degree of incompleteness given $\mathbf{z}^{BT}(\mathbf{x})$ normalized between 0 and 1.

$$w(a, b) = \begin{cases} 1 & a = b \\ \frac{(t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})))}{\sum_b \exp(t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})))} & else. \end{cases}$$

Objective 3: This objective is similar to the previous two, but it simply removes the denominator of w to make its sign depend only on the numerator.

$$\mathcal{L}_{contrast} = \frac{-1}{|\mathbf{S}|^2} \sum_a \sum_b w(a, b) (\lambda_1 \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) + (1 - \lambda_1) \log (1 - \log t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt}))),$$

where $0 < \lambda_1 < 1$,

$$t'_*(\mathbf{z}^{BT}(\mathbf{x})) = \frac{1}{T} \sum_{s=0}^T s \cdot p_\psi(t'|\mathbf{z}^{BT}(\mathbf{x}))$$

is the expected degree of incompleteness given $\mathbf{z}^{BT}(\mathbf{x})$ normalized between 0 and 1.

$$w(a, b) = \begin{cases} 1 & a = b \\ t'_*(\mathbf{z}^{BT}(\mathbf{x}_b^{txt_{inc}})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})) & else. \end{cases}$$

Objective 4: This objective is the same as the proposed one except that it does not consider cross-pair comparison. The actual t'_* for the embeddings becomes either close to 0 or 1. This is because when the contrastive loss is minimized, the margin between a given note and its incomplete note is maximized. By considering cross-pairs, the proposed contrastive loss has a buffer to avoid extreme values of t'_* .

$$\mathcal{L}_{contrast} = \frac{1}{|\mathbf{S}|} \sum_a t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt})) - t'_*(\mathbf{z}^{BT}(\mathbf{x}_a^{txt_{inc}})),$$

	τ	ρ
Objective 1	0.398 _{0.012}	0.410 _{0.008}
Objective 2	0.405 _{0.014}	0.419 _{0.006}
Objective 3	0.203 _{0.009}	0.227 _{0.006}
Objective 4	0.178 _{0.013}	0.182 _{0.005}
Proposed	0.915 _{0.010}	0.928 _{0.008}

Table 6: Compare the ranking performance of estimated t'_* learned by different objectives. The dataset for evaluation is the synthetic notes with deletion of 2, 4, and 6 total number of sentences progressively.

C Baseline

For evaluation, we consider the following baselines which are proposed for prediction based on multimodal data.

1. MT (Lyu et al., 2023) discovers important structured EHR features with important words in clinical notes.
2. UTDE (Zhang et al., 2023b) tackles temporal irregularity of multimodal data through a time attention mechanism and model irregular time series by regular interpolation embeddings.
3. MCP (King et al., 2023) extracts meaningful patterns by self-supervised pretraining for the alignment of multimodal data.
4. FuseMoE (Han et al., 2024) is designed to deal with a diverse number of modalities with temporal irregularity and missing modalities through an innovative gating function.
5. AUTOFM (Cui et al., 2024) automates the process of model design for mining EHR data with a novel neural architecture search.
6. MoE-Retriever (Yun et al., 2025) address the problem of missing modalities by considering intra- and inter-modal contexts by leveraging context-aware embeddings.
7. CTPD (Wang et al., 2025) identifies critical temporal patterns across modalities.

D Datasets

D.1 Real-World Data

We prepare three datasets from MIMIC-III (Johnson et al., 2016), each of which consists of tabular time series, a sequence of notes, and a class label. MIMIC-III is a public dataset and free to use upon successful registration at <https://physionet.org/>. The time series in all datasets share the same patient features. Examples include vital signs (heart rate, respiratory rate, etc.) and lab test (white blood cell count, platelet, albumin, etc.). Patients of high feature-missingness are excluded. The missing values are imputed by patient-specific median if available, otherwise, population median. Note that the median is calculated only based on the data within an observation window (e.g., 48 hr) to avoid information leakage. To extract nursing notes, we select documentation of category “Nursing/other”. To

ensure the documentation is written by registered nurses, we select those with a caregiver label of “RN”. The minimum number of characters of inclusion is 100 for ensuring the documentation content is non-trivial. Also, we avoid selecting documentation containing key words such as “addendum” and “management note”, which indicate that it does not belong to a nursing note. The feature dimension for time series is 37. To create $x^{txt,inc}$, the incomplete text, for our contrastive learning, we randomly delete 1-12 sentences from a given nursing note.

For the task in-hospital mortality (IHM), we focus on the first admission to ICUs for the sepsis cohort. We set the observation window to be 48 hours, which means we only select both time series and notes within the first 48 hours of ICU stay. Patients die within the observation window are excluded. The label indicates whether the patient will die by the end of ICU stay. The total number of patients in this dataset is 10,973, among which 11.2% are positive cases.

For the task length of stay (LOS), we focus on the first admission to ICUs for the sepsis cohort. The observation window is 48 hours. The labels indicates whether a patient will be discharged in the following 48 hours. Patients who die within the observation window are excluded. The total number of patients in this dataset is 11,472, among which 39.4% are positive cases.

For the task septic shock classification (SSC), we focus on the first admission to ICUs for the sepsis cohort, who stay in ICUs less than 72 hours (the number of patients are drop almost by 50% if the length of stay is less than 48 hours), which is the observation window. The label indicates whether a patient is classified as septic shock. The total number of patients in this dataset is 5,966, among which 10.3% are positive cases.

D.2 Synthetic Data Generation

We use DeepSeek-R1 to generate 10k synthetic nursing notes by using the following prompt:

According to the following basic information of a septic patient, generate a synthetic nursing progress note. Make sure omit the confidential information including people’s names, date, and room number. Only include these aspects, which are Neuro, Cardiac, Resp, skin, GU/GI, ID, assessment, and plan. Provide as much details as you can. Total token limit is 512. Output the note only. The basic information is: {patient info}.

In this prompt we specify the sections for a com-

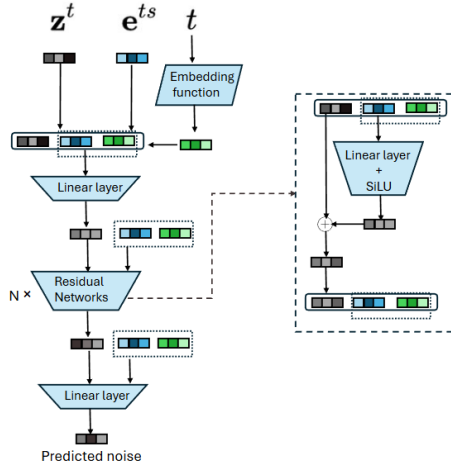


Figure 5: This figure shows the architecture of our denoiser ϵ_θ . The residual networks adopts the residual connection (He et al., 2016).

plete note and instruct the LLM to provide details as much as possible for the purpose of completeness. The *{patient info}* is a structural statement like:

Name: {name}, Age: {Age}, Gender: {Gender}, Blood Type: {Blood Type}, Date of Admission: {Date of Admission}, Doctor: {Doctor}, Hospital: {Hospital}, Insurance: {Insurance}, Billing Amount: {Billing Amount}, Room Number: {Room Number}, Admission: {Admission}, Discharge: {Discharge}, Medication: {Medication}, Test Results: {Test Results}.

The information in a bracket is from a synthetic dataset of basic patient information (Patil).

E Model Training

E.1 Denoiser Architecture

The architecture of the denoiser is shown in Fig. 5.

E.2 Hyperparameter Fine-tuning

Throughout all experiments, the maximum number of training epoch is 2,000 with early-stop tolerance 12. For each set of hyperparameters, we run a method for 5 times and then calculate the statistics of the metrics with Python package scikit-learn. At each time, data are randomly split into training, validation, and test sets with a split ratio of 6:2:2, whose batch size is 128 in all experiments. The optimal metrics are reported in the paper. We use short-hand notations to denote some common hyperparameter names: “lr” = learning rate, “h” = size of a hidden layer, “L” = number of hidden layers, “emb dim” = dimension of embeddings, “time emb” = dimension of time embeddings, “proj dim” = the target output dimension of projection layers,

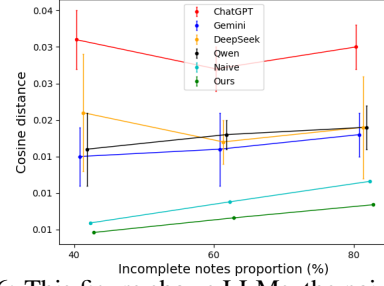


Figure 6: This figure shows LLMs, the naive baseline, and our method on the recovery of [CLS] embeddings on the synthetic data.

“heads” = the number of attention heads, “expert” = the number of experts, “k” = top k, “λ” = a coefficient. If a hyperparameter is not specified for a method, it is either because this hyperparameter is not applicable or the same choice in the original public source code is adopted. Tables 7, 8, and 9 include the choices of hyperparameters.

The proposed method CODE was trained on a single NVIDIA H200 GPU (140GB), with 1 CPU (Intel Xeon Platinum 8480+), 258GB of system memory, and 30GB of disk. The operating system was Linux 8.9, running Python 3.11. The environment was based on PyTorch 2.10.0 and CUDA 12.7. We set gradient accumulation to 4. The module of $p_\psi(\cdot)$ is first pretrained until convergence and is frozen, after which MoE-Retriever modules are pretrained with 10 epochs. Then, all modules are trained until convergence. CODE has around 21M parameters and roughly takes 14 GPU hours to train for one time.

E.3 LLM Prompts

Given the observed synthetic nursing notes, LLMs are instructed to output the potential complete text. To do so, we use the following prompt:

The following nursing progress note may or may not have sentences deleted. A progress note should include sections Neuro, Cardiac, Resp, skin, GU/GI, ID, assessment, and plan. Add sentences to the given note to make it complete if necessary. Output the note only.

{nursing note}.

The prompt specified the sections to be included in a complete note and also informs the LLM the missing pattern is sentence deletion.

Hyperparameters	MT	UTDE	FuseMoE
lr	{0.001,0.0001,0.00001}	-	-
lr for time series	-	{0.001,0.0001,0.00001}	{0.001,0.0001,0.00001}
lr for notes	-	{0.001}	{0.001}
L	-	{2,3}	{2,3}
emb dim	{256, 128, 64}	{128, 160}	{128, 160}
proj dim	{512, 256, 128}	-	-
heads	-	{8}	{8}

Table 7: Hyperparameters of MT, UTDE, and FuseMoE.

Hyperparameters	MCP	MoE-Retriever	AUTOFM
lr	{0.001,0.0001,0.00001}	{0.001,0.0001,0.00001}	{0.001,0.0001,0.00001}
h	{64, 128, 256}	-	-
L	-	{2,3}	-
emb dim	-	-	{256, 128, 64}
expert	-	{8,12,16,32}	-
k	-	{2,4}	-
λ	-	-	{0.05, 0.1, 0.2}

Table 8: Hyperparameters of MCP, MoE-Retriever, and AUTOFM.

Hyperparameters	CTPD	CODE
lr	{0.001,0.0001,0.00001}	{0.001,0.0001,0.00001}
L	-	{2,3}
emb dim	{128, 160}	{128, 160}
time emb	{64,128}	{128, 160}
expert	-	{8,12,16,32}
k	-	{2,4}
λ_0	-	{0.2,0.4,0.5,0.6,0.8}
λ_1	-	{0.01,0.2,0.4,0.6,0.8,0.99}
λ_2	-	{0.2,0.4,0.6,0.8}
T	-	{40}

Table 9: Hyperparameters of CTPD and CODE.

F Case Study

Here we show one of five cases where “removed sentences” are quoted and followed by the comments of our clinician. For the patient in this case, there two nursing notes. In the first note:

1. “*pt unresponsive, does not move extremities at all*” – this sentence provides the neurological assessment, which is important in outcome prediction (important fact).
2. “*elimination: incontinent formed brown stool with some brb, stool guaiac+, also has menses*” – the incontinent represents a total loss of consciousness, which is important in outcome prediction (important fact). The color of the

stool or the presence of menses is not important (uncertain).

3. “*chest tube has air leak, some small amt crepitus still present in left chest*” – the air leak in the chest tube represents a persistent hole in the left lung, which could be for outcome prediction (important fact).

In the second note:

1. “*ebrile, tachycardic, hypertensive, tachypneic with significant metabolic alkalosis.*” – vital signs that are useful but may not be very predictive (important fact).
2. “*brief episode of hypotension (60-70/) on return from ct when positioned with left side*

down” – an important progress that signifies instability of patient’s condition (important fact).

3. “suctioned infrequently for thick, tan, blood tinged secretion” – an observation that implies significant lung injury (important fact).
4. “skin: intact, starting to third space” – this is a factual observation, but probably not important to the outcome (uncertain).

G Additional Results

[CLS] of	Avg. t'_*
CODE	0.111 _{0.003}
Given Notes	0.182 _{0.007}
Incomplete Notes	0.248 _{0.005}

Table 10: Compare the average t'_* of the [CLS], which is generated by the proposed method from the given [CLS], with that of the [CLS] of both given and incomplete notes.

G.1 ClinicalBERT as Encoder

Table 11 shows the results for CODE on real-world data using ClinicalBERT (Huang et al., 2020) as the encoder. Also, Figure 7 shows the robustness of our method against missing information.

G.2 Ablation Study

Similar patterns of gains can be observed in Tables 12 & 13 referring to the ablation study on LOS and SSC.

G.3 Ranking and Similarity Tests

Figure 6 shows that the semantics of the LLM generated complete note is even deviated from the ground truth as a comparison to the observed notes. Table 10 shows that the estimated t'_* can represent degree of completeness that the average value of incomplete notes is higher than given notes. The low t'_* of the embeddings generated by our model indicates our model improves the semantic completeness of input embeddings.

G.4 Omission Depending on Clinical Significance

Regarding the contrastive learning of the proposed method, incomplete notes are created by random deletion of any sentence of given notes. We investigate random deletion of sentences classified as “stable routine findings” or “unusual findings”

by an LLM. Table 14 considers omissions only on “stable routine findings” and the results are suboptimal to the proposed one at both IHM and LOS tasks. Next, we consider omissions only on “unusual findings”, whose results shown in Table 15 are still suboptimal to the proposed method.

G.5 Simple Baselines

We investigate more choices of model architecture using simple baselines. Specifically, we consider denoising autoencoder (Bengio et al., 2013) and MoE-Retriever with the augmented data of incomplete notes. Also, we consider single modality for time-series only and text only, for which the representations for downstream tasks are learned by self-attention. Their performance on IHM is shown in Table 16.

G.6 Sample Notes

Real Notes Ranking Figure 8 contains examples of top, middle, and bottom notes ranked by $p_\psi(t' = 0|\mathbf{z})$ for IHM dataset. The subsection headings are highlighted in blue, for which we can see the top notes have descriptions more detailed than middle notes. Bottom notes clearly have lowest degree of completeness.

LLM Generation Figures 9 & 10 are examples of using zero-shot methods to provide missing information of an incomplete note. The ground truth is shown where missing information is highlighted in blue. LLMs are prompted to generate the potentially complete note, which only see the ground truth without blue text. The text in red refers hallucination of LLM responses. We can see that ChatGPT has a serious hallucination problem. We can also see that LLMs sometimes are not able to detect missing information, to which they have no response. When responding to the missing information such as sections “neuro” or “cardiac”, LLM can suffer hallucination.

Table 11: Performance comparison on three prediction tasks. Metrics are reported as Mean_{Std}. ClinicalBERT is used as the encoder.

		MT	UTDE	FuseMoE	MCP	MoE-Retriever	AUTOFM	CTPD	CODE (ours)
IHM	F1	0.372 _{0.060}	0.494 _{0.038}	0.508 _{0.016}	0.326 _{0.029}	0.514 _{0.019}	0.505 _{0.017}	0.518 _{0.019}	0.528_{0.023}
	AUPRC	0.399 _{0.027}	0.578 _{0.021}	0.566 _{0.022}	0.353 _{0.017}	0.581 _{0.020}	0.565 _{0.020}	0.585 _{0.010}	0.596_{0.017}
	AUROC	0.834 _{0.011}	0.902 _{0.007}	0.900 _{0.007}	0.788 _{0.021}	0.903 _{0.008}	0.898 _{0.007}	0.908 _{0.005}	0.910_{0.007}
LOS	F1	0.667 _{0.049}	0.744 _{0.013}	0.737 _{0.014}	0.628 _{0.034}	0.750 _{0.014}	0.741 _{0.016}	0.748 _{0.009}	0.763_{0.010}
	AUPRC	0.747 _{0.020}	0.799 _{0.015}	0.790 _{0.019}	0.608 _{0.031}	0.803 _{0.013}	0.792 _{0.015}	0.785 _{0.014}	0.809_{0.011}
	AUROC	0.827 _{0.009}	0.871 _{0.006}	0.866 _{0.009}	0.724 _{0.029}	0.868 _{0.005}	0.869 _{0.005}	0.867 _{0.005}	0.879_{0.006}
SSC	F1	0.392 _{0.026}	0.393 _{0.028}	0.370 _{0.026}	0.208 _{0.033}	0.399 _{0.036}	0.288 _{0.067}	0.392 _{0.039}	0.411_{0.028}
	AUPRC	0.383 _{0.028}	0.431 _{0.035}	0.416 _{0.045}	0.162 _{0.009}	0.428 _{0.027}	0.367 _{0.035}	0.411 _{0.043}	0.445_{0.021}
	AUROC	0.803 _{0.004}	0.847 _{0.013}	0.837 _{0.020}	0.602 _{0.023}	0.843 _{0.016}	0.819 _{0.021}	0.829 _{0.022}	0.859_{0.017}

Table 12: Ablation study. LOS.

	$+\mathcal{L}_{denoise}$	$+\mathcal{L}_{contrast}$	$+\mathcal{L}_{complete}$	F1	AUPRC	AUROC
	✗	✗	✗	0.621 _{0.0012}	0.696 _{0.012}	0.779 _{0.009}
	✓	✗	✗	0.655 _{0.009}	0.727 _{0.015}	0.806 _{0.007}
	✗	✓	✗	0.625 _{0.013}	0.697 _{0.018}	0.783 _{0.010}
	✗	✗	✓	0.681 _{0.014}	0.752 _{0.013}	0.828 _{0.009}
	✓	✓	✗	0.709 _{0.008}	0.772 _{0.012}	0.843 _{0.008}
	✓	✗	✓	0.697 _{0.008}	0.766 _{0.017}	0.838 _{0.006}
	✗	✓	✓	0.727 _{0.012}	0.785 _{0.013}	0.847 _{0.009}
	✓(no $\mathcal{L}_{0:t'}$)	✓	✓	0.742 _{0.011}	0.796 _{0.011}	0.873 _{0.008}
	✓	✓	✓	0.757 _{0.012}	0.818 _{0.011}	0.876 _{0.006}

Table 13: Ablation study. SSC.

	$+\mathcal{L}_{denoise}$	$+\mathcal{L}_{contrast}$	$+\mathcal{L}_{complete}$	F1	AUPRC	AUROC
	✗	✗	✗	0.349 _{0.041}	0.385 _{0.031}	0.772 _{0.015}
	✓	✗	✗	0.368 _{0.047}	0.402 _{0.032}	0.792 _{0.019}
	✗	✓	✗	0.352 _{0.046}	0.389 _{0.027}	0.773 _{0.014}
	✗	✗	✓	0.381 _{0.053}	0.412 _{0.023}	0.804 _{0.017}
	✓	✓	✗	0.396 _{0.045}	0.422 _{0.036}	0.816 _{0.015}
	✓	✗	✓	0.394 _{0.057}	0.416 _{0.030}	0.812 _{0.019}
	✗	✓	✓	0.405 _{0.052}	0.427 _{0.027}	0.829 _{0.020}
	✓(no $\mathcal{L}_{0:t'}$)	✓	✓	0.404 _{0.043}	0.436 _{0.025}	0.842 _{0.017}
	✓	✓	✓	0.426 _{0.058}	0.453 _{0.037}	0.860 _{0.017}

	F1	AUPRC	AUROC
IHM	0.541 _{0.018}	0.594 _{0.018}	0.910 _{0.004}
LOS	0.752 _{0.015}	0.812 _{0.010}	0.876 _{0.005}

Table 14: CODE evaluated with random omission on “stable routine findings” only.

	F1	AUPRC	AUROC
IHM	0.536 _{0.014}	0.593 _{0.016}	0.909 _{0.005}
LOS	0.748 _{0.016}	0.810 _{0.013}	0.876 _{0.004}

Table 15: CODE evaluated with random omission on “unusual findings” only.

	F1	AUPRC	AUROC
Denoising autoencoder	0.388 _{0.020}	0.424 _{0.017}	0.846 _{0.005}
MoE-Retrivier (aug. data)	0.511 _{0.022}	0.576 _{0.014}	0.903 _{0.006}
Text only	0.294 _{0.013}	0.336 _{0.019}	0.771 _{0.003}
Time-series only	0.347 _{0.014}	0.362 _{0.023}	0.760 _{0.002}

Table 16: Simple baselines evaluated on IHM.

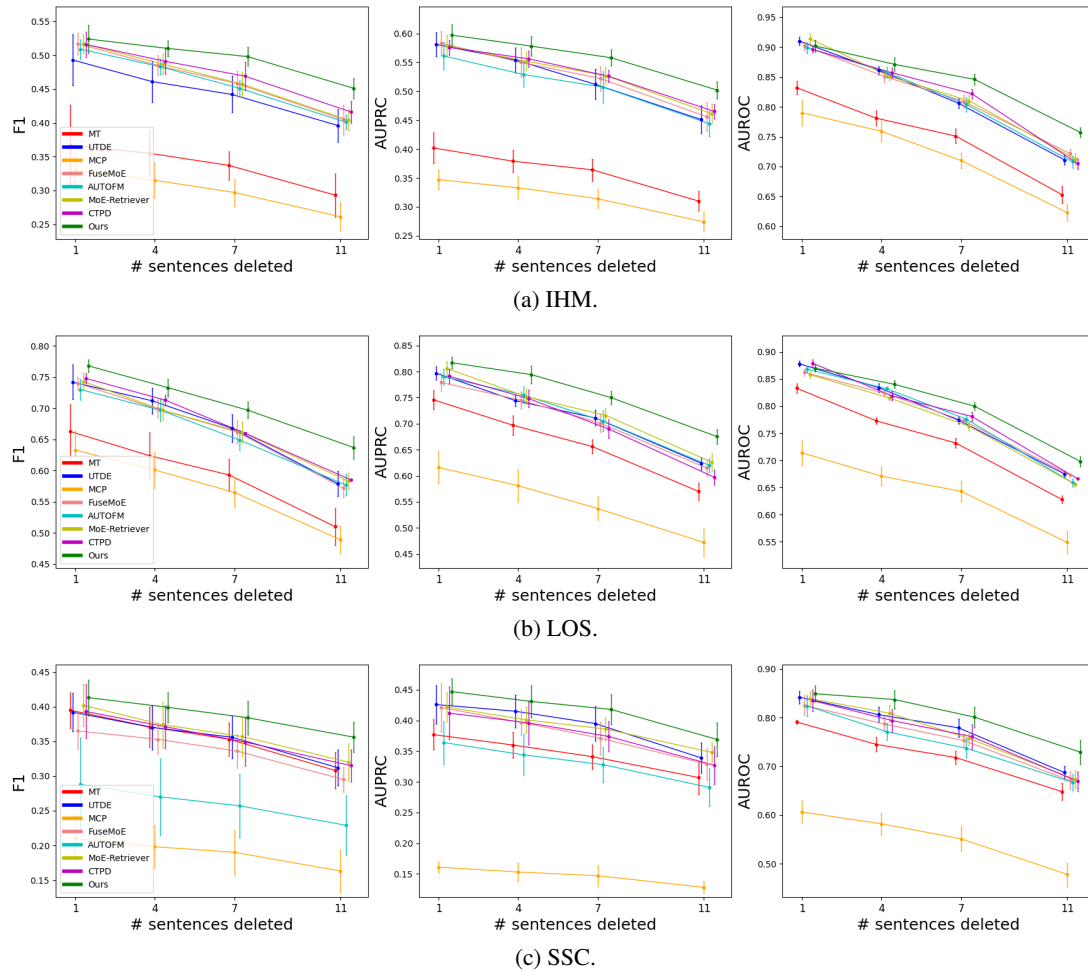


Figure 7: Performance comparison on in-hospital mortality (a), length of stay (b), and septic shock classification (c) with different sentences deleted from nursing notes. Use ClinicalBERT as the encoder.

Ranked by $p_{\psi}(t' = 0|\mathbf{z})$

Top 2 Notes	
tsicu npn (2300-0700) review of systems: neuro : continued poor neuro exam. nonresponsive to any stimuli. pupils 3mm bilaterally and nr. absent gag, cough, and corneal reflexes. no movement noted. does overbreath venttachypneic and discordant pattern. propofol and fentanyl drips started yesterday in attempts to slow down resp rate. appears to have minimal effectivenesspt still tachypneic and discordant even after boluses. j collar on. t1s has not been cleared per dr. dilantin tid on board. cv : nrs with rare ectopy noted. hr 70's. initially upon assessment pt on labetalol drip due to elevated hr and hypertension. weaned off by 0200 as bp kept dropping. has sustained 100-120's/60-70's off labetalol. resp : extremely tachypneic and discordant breathing pattern. initially on cpap 5/5/40% with rr in low 40's, tv's 250-350, and sats >95%. deadspace had been added per rt prior to this nurse assuming care in hopes to help pt retain some co spoke with dr about rr and tv's. no action desired at that time. am abg at 0200 12, 102, 103.0, notified dr. deadspace removed per rt. changed mode to inv 500x18, 5/5/40%. abg after change 35/47/134/0/ rr down to 30's. lung fields coarse. suctioning scant white, thin secretions. gi : abd soft with hypoactive bs. ogt to lws with small amts of dark red secretions. pepid coverage on board. gu : adequate but minimal urine output. ivf at 70cc/hr. mag repleted. heme...hct stable at heparin subq. pneumo boots on. id : wbc cool catheter in place to cool pt with target temp 36c. working well. endo : coverage per riss. skin : no skin issues. backside intact. small abrasions to anterior neck. social : wife burying her mother today. not sure if she will be in to visit or not. no contact overnight. plan : monitor neuro status. eeg and mri ordered for today. monitor resp status. pulmonary toilet. replete lytes. ???start planning for family meeting after today	rn-11pm-7am: neuro : intubated on propofol@ 60mcg/kg/min.change in vs with repositioning.morphine 2-4mg iv given q2-3 hrs for comfort.pupils 2 sluggish.unable to do daily wakeu/hemodyn. unstable. cv : sr 83- rare pvc's.av paced @82 at beginning of shift.rate down to maintaining own rate and bp.swan changed at beginning of shift(leaking).good placement per ccray.cco svo@not accurate.positional.mix venous 45/53/epinephrine drip increased from 015 to 03 with good effect.ci>hypertensive. bp 140-150's at times.nitro @3mcg/kg/min.hydralazine 10mg given with good effect.bp 110-130's per aline.2 units of rbc given.repeat hct pulses per doppler.la/ue cool.obese with generalized edema.weight up.net +/- 3L(baseline pre-op very hypertensive 180- 200 syst.) resp : intubated.sat 92-94 %.peep increased from 8 to rr 24-using assessor muscles.propofol to max.ccray should pulmonary edema.po2 dropped from 81 to lasix 20mg given.poor effect.lasix 40mg iv repeated.o@ sat improved.bun/creat wnl.small yellow/tan secretions.ls diminished bil.ct @20 cm suction draining sero-sang moderate amt. gi/gu : foley to gravity.poor u/o.lasix iv given.bun/creat wnl.no bs.abdomen soft/obese.og minimal bilious.npo. endo : insulin drip per protocol.bs<1 hr fs. id : wbc afebrile.waco post/op. skin : cool and dry.edema +++ obese.no breakdown.coccyx intact.criticalid cream applied.tuned q2 hr.echymotic areas(old iv sites?)left acewrap intact. plan : keep intubated until hemodyn.stable.monitor labs.pain management.tirate nitro/epi/protocol md
Middle 2 Notes	
11p-7a pod 1 cabgx4, avr, maze requiring levo, epi and liters of fluid for hemodynamic stability. neuro : sedated on propofol. nodded to pain when light. medicated with morphine 2mg x2 with effect. perl on sluggish side. cv : apaced, initially at 88 but decreased to 70 after echo as reported lv would benefit from increased filling time. bp labile 80- 160/50- co low initially 8, increasing to 7 through night. ci 8- pt remains on epi at .025mcg and levo weaned to .09mcg. 1 unit prbc given overnight, along with multiple liters of fluid to keep bp and index up. +pp. lactic acid 8 (9). resp : orally intubated ac 60%. abg improved, about normal. lungs clear. sats 98%. svO2 improved as well- 58% this am, up from does drop with turn. minimal chest tube output. gi/gu : abd obese, soft, no bs. ogt to lows draining bilious. cont on insulin gtt with bs skin : small bruise on buttocks. bed rotation on. see flowheet for other. social : son will be in today. plan : cont to wean levo as pt tolerates, monitor hemodynamics closely. monitor resp status, begin vent wean when pt will tolerate. check with team about plan to wean epi. monitor labs,treat prns when bs allows. pain control.	nursing note: neuro : arouses to verbal stimulation, moves ues. agitated at times, given fentanyl for c/o's pain, begun on low dose pprf gtt. perla 3mm and brisk. does nod to 7's. resp : lung sounds coarse, dzim to bases. suctioned infrequently for blood-tinged secretions. changed to pcv, atgs acceptable. biting down on ett at times, dropping sats and vs, bite block in place. cv : hypothermic @95, warm blankets on. sr w/occ pvc's in 80's. cvp 17- pads 20- pa catheter advanced slightly by dr. to 51 cm. given platelets for plt count 36, oozing from mult. line sites and mouth/gums. +anasarca throughout. echo done. gi : abdomen large, +ascites. +liquid black tarry stool, sent for cdiff and cxs. - bs. us done of liver and of les. gu : foley patent minimal amount cloudy amber urine. ? begin cvvhd. endo : glucose stable, given ssri prn. skin : breakdown to coccyx noted, tegaderm in place, no drainage. leaking ascites from mult. sites. social : pt's mother updated by team, will be in this evening. ap : hemodynamically stable pt. w/esld and likely hepatorenal syndrome undergoing ? tx work-up. continue to monitor closely, ? cvvhd, frequent labs
Bottom 2 Notes	
1 infant with potential sepsis 2 rds 3 fen revisions to pathway: 1 infant with potential sepsis; added start date: 2 rds; added start date: 3 fen; added start date:	1 altered resp. status 2 fen 3 growth and development 4 infant with potential sepsis revisions to pathway: 1 altered resp. status; added start date: 2 fen; added start date: 3 growth and development; added start date: 4 infant with potential sepsis; added start date:

Figure 8: Two top, middle, and bottom nursing notes for IHM, which are ranked by the estimated probability of being complete. Enlarge to see more clearly.

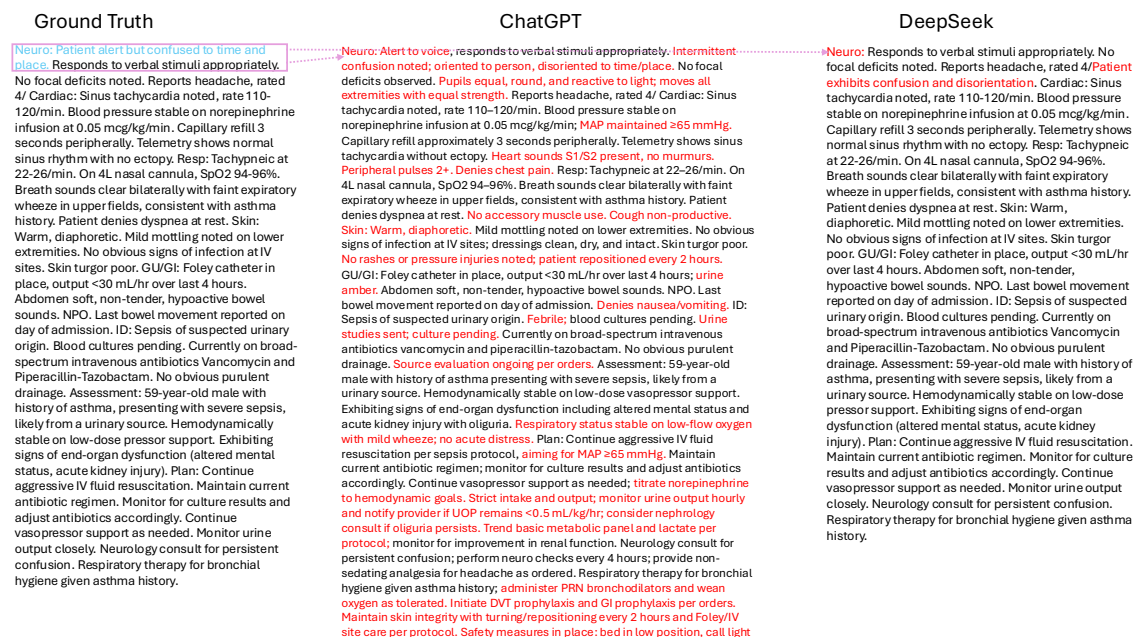


Figure 9: This figure shows the example responses from both DeepSeek and ChatGPT for the same observed note. The observed note to LLMs is the ground truth with blue text removed.

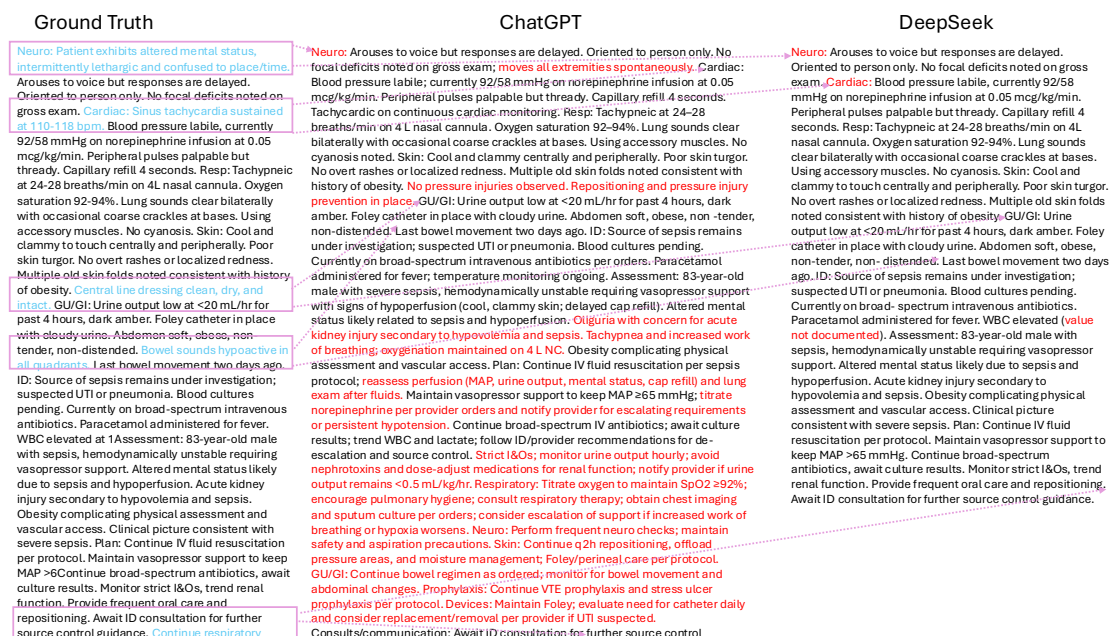


Figure 10: This figure shows the example responses from both DeepSeek and ChatGPT for the same observed note. The observed note to LLMs is the ground truth with blue text removed.