Addressing Asynchronicity in Clinical Multimodal Fusion via Individualized Chest X-ray Generation

Wenfang Yao[1]*, **Chen Liu** [1,3]*, **Kejing Yin** [2]†, **William K. Cheung**[2], **Jing Qin**[1]
[1]School of Nursing, The Hong Kong Polytechnic University
[2]Department of Computer Science, Hong Kong Baptist University
[3]School of Software Engineering, South China University of Technology

Abstract

Integrating multi-modal clinical data, such as electronic health records (EHR) and chest X-ray images (CXR), is particularly beneficial for clinical prediction tasks. However, in a temporal setting, multi-modal data are often inherently asynchronous. EHR can be continuously collected but CXR is generally taken with a much longer interval due to its high cost and radiation dose. When clinical prediction is needed, the last available CXR image might have been outdated, leading to suboptimal predictions. To address this challenge, we propose DDL-CXR, a method that dynamically generates an up-to-date latent representation of the individualized CXR images. Our approach leverages latent diffusion models for patient-specific generation strategically conditioned on a previous CXR image and EHR time series, providing information regarding anatomical structures and disease progressions, respectively. In this way, the interaction across modalities could be better captured by the latent CXR generation process, ultimately improving the prediction performance. Experiments using MIMIC datasets show that the proposed model could effectively address asynchronicity in multimodal fusion and consistently outperform existing methods.

1 Introduction

Clinical data in modern healthcare is documented through various complementary modalities [1, 2]. Electronic health records (EHRs), for instance, systematically record the progression of diseases over time, including medical histories, laboratory test results, and treatment outcomes [3–6]. In parallel, medical imaging, such as chest X-rays (CXRs), is valuable for providing visual insights into the patient's internal anatomy, organ functions, and potential abnormalities [7]. Recent studies have shown that strategic integration of multimodal clinical data could lead to improved performance for clinical predictions compared to relying solely on uni-modal data [8–13].

Despite the promising results obtained, the inherent asynchronicity of multimodal clinical data still hinders effective integration. Take the intensive care unit (ICU) setting as an example, patients are subject to continuous monitoring systems that capture vital signs, including heart rate, blood pressure, and oxygen saturation, with this information being routinely recorded in the EHR [14, 15]. On the other hand, CXRs are captured only on an as-needed basis and often as less as possible, due to limitations of radiation dose and resources [16]. However, patients admitted to ICU are in life-threatening conditions, which means their medical status is prone to rapid changes and highly time-sensitive [17]. In the MIMIC-CXR dataset [18], it is observed that among patients with positive disease findings in their CXR, over 70% of subsequent CXR images — taken within a median interval

---

*These authors contributed equally.
†Correspondence to: Kejing Yin <cskjyin@comp.hkbu.edu.hk>

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

| (a) Initial Chest X-ray | (b) CXR taken after 34 hours | (c) Generated by DDL-CXR |

Figure 1: A real ICU patient with rapid CXR changes. (a) *Initial radiology findings*: Low lung volumes but lungs are clear of consolidation or pulmonary vascular congestion. No acute cardiopulmonary process. (b) *Radiology findings after 34 hours*: Severe relatively symmetric **bilateral pulmonary consolidation**. (c) CXR generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 34 hours. Clear signs of bilateral pulmonary consolidation can be seen from the generated image. The visualization shows that DDL-CXR **could generate updated CXR images that respect the anatomical structure of the patient and reflect the disease progression.**

of less than 24 hours — exhibit changes in CXR findings. This implies that when a clinical prediction is needed, CXRs captured even only a few hours ago could have become outdated, especially for ICU patients who commonly have respiratory, cardiac, infectious, and traumatic conditions [19]. Fig. 1 shows such an example of a real patient in the MIMIC dataset.

Motivation   Existing works adopt the "carry-forward" approach, i.e., using the last CXR image available for downstream prediction tasks [20, 10]. This strategy ignores the potential rapid changes between the prediction time and the time of the last CXR image taken and thus inevitably leads to suboptimal prediction performance. On the contrary, we hypothesize that generating an updated CXR image at the prediction time could mitigate the asynchronicity problem and enhance the prediction accuracy. Nevertheless, generating patient-specific CXR images presents unique challenges. While multimodal generation has been explored extensively in various fields, these methods are not readily adaptable for generating individualized CXR images. In domains such as text-to-audio [21] or text-to-image generation [22], the attributes that need to be controlled (e.g., painting style) can be explicitly defined in input modalities (e.g., the text prompt). However, in the clinical context, explicit descriptions of a patient's anatomical structures, organ functions, and disease progression, which are highly specific to individual patients and critical for downstream prediction, are not directly available.

Contribution   To tackle the aforementioned challenge, we propose *Diffusion-based Dynamic Latent Chest X-ray Image Generation* (DDL-CXR)[3], which utilizes a tailored latent diffusion model (LDM) [22] to generate individualized CXR images for clinical prediction. Specifically, DDL-CXR learns to generate representations in a latent space encoded by a variational auto-encoder (VAE). To incorporate detailed information about the patient's anatomical structure and organ specifics, we use a previous CXR image from the same patient as the reference image. To generate latent representations that align with the disease progression, we use a Transformer model [23] to encode the irregular EHR data spanning from the reference CXR to the prediction time. To further capture the implicit interactions between EHR and CXR, we use the encoded EHR representation to predict the labels of abnormality finding of the target image. To force the LDM to capture the disease course in the EHR data, we explore a contrastive learning approach for training the LDM. The generated up-to-date latent CXR is later fused with historical data for downstream clinical prediction.

We summarize our contributions as follows:

- To our knowledge, DDL-CXR is the first work to generate an updated individual CXR image to improve clinical multimodal fusion, thereby alleviating the asynchronicity between EHR and CXR.
- We propose a contrastive learning approach for the LDM training to enable the disease course in EHR to be captured and utilized by the LDM.
- Experiments show that DDL-CXR outperforms existing methods in both multi-modal clinical prediction and individual CXR generation.

---

[3]The code is available at https://github.com/Chenliu-svg/DDL-CXR.

2

2 Related Work

Clinical multi-modal fusion   Integrating multi-modal clinical data has shown beneficial for various clinical prediction tasks [24], including COVID-19 prediction [25], pulmonary embolism diagnosis [8, 26], AD diagnosis [9] and X-ray image abnormality detection [27].

Different strategies have been proposed to facilitate the fusion of multi-modal clinical data [2, 28]. Hayat et al. [10] adopts feature-level fusion with an LSTM layer, while Zhang et al. [29] utilizes a modality-correlated encoder to capture long-range dependencies across modalities. Zhang et al. [30] and Lee et al. [12] incorporate modality type embedding into the self-attention to capture the interaction. Despite the effort, existing methods for multi-modal fusion are driven only by downstream predictions. How to capture the more fundamental interaction between different data modalities remains an open challenge.

In the temporal setting, asynchronicity presents another major challenge. Unlike the settings of medical images and radiology reports [31], which are naturally aligned in time, EHR and CXR are often highly asynchronous, bringing extra difficulties to information integration. "Carry-forward" is a common strategy adopted, where the last available data from different modalities are used [10, 13]. Lee et al. [12] and Zhang et al. [17] also adopt this approach while modeling the time information of the last available data.

Conditional latent diffusion models   The diffusion model is one of the state-of-the-art generative models [32, 33] that has found important applications in areas such as image generation [34], sound generation [35], joint audio and video generation [36], and tabular data generation [37]. To reduce the computational cost, LDM [22] proposes to train diffusion models on a latent space encoded via pre-trained VAE, thus improving training and sampling efficiency as well as preserving generation quality. It also incorporates an attention mechanism into its underlying neural backbone to allow more flexible conditioning.

Based on LDM, multi-modal generation models have been developed using priors obtained from large-scale contrastive pre-training, e.g., contrastive-image pairs for text-to-image generation [38] and contrastive language-audio pairs for text-to-audio generation [21]. However, it is infeasible to apply this method to clinical settings since many clinical data modalities, e.g., CXR and EHR, capture different aspects of patients and cannot be semantically aligned like the image and caption pairs as in CLIP.

In clinical settings, LDM-based models are developed for brain MRI image generation, conditioned on age, sex, brain structure volumes [39], and a subset of MRI slices [40]. For CXR image generation, Packhäuser et al. [41] adopts a thoracic abnormality classifier-aided LDM to generate anonymous CXR images for privacy-protected data generation. Weber et al. [42] utilizes pathology labels, radiological reports, and radiologists' annotations for synthesizing customized CXR images. Gu et al. [43] explores counterfactual generation for CXR using information from imaging reports. Generating individual CXR images that reflect disease courses in EHR and applying them to medical predictions remains an open challenge.

3 `DDL-CXR`: The Proposed Method

In this work, we focus on improving multimodal clinical predictions by generating latent CXR images that are in line with patient conditions at prediction time. The generation process also works as a fusion module that captures the cross-modal interaction between EHR and CXR. The overview of `DDL-CXR` is depicted in Fig. 2. It consists of two stages: the LDM stage and the prediction stage. In the LDM stage, we use the consecutive image pairs to train an LDM that generates representations within a latent space encoded by a variational autoencoder (VAE). To generate patient-specific CXRs, an earlier CXR image of the same patient is used as a reference to capture the anatomical structure, and the EHR time series between the consecutive image pairs is used to capture the disease progression. In the prediction stage, conditioned on this composite information, `DDL-CXR` generates updated and informative CXR representations at prediction time, which are subsequently fused with available EHR data as well as the previous CXR image for downstream prediction tasks.

3

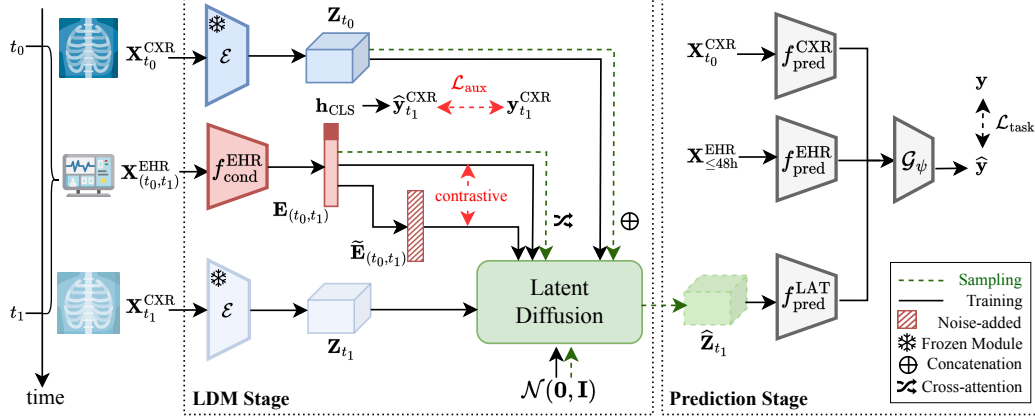



Figure 2: The overview of the proposed framework DDL-CXR. It consists of two stages. The **LDM stage** learns to generate an individualized up-to-date latent CXR at time t_1, $\hat{\mathbf{Z}}_{t_1}$, to address asynchronicity by conditioning on a previous CXR image taken at time t_0, $\mathbf{X}_{t_0}^{\text{CXR}}$, which provides the anatomical structure of the patient, as well as EHR data between t_0 and t_1, $\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}$, that provides information on disease progression. A contrastive loss and auxiliary loss are enforced for better EHR information integration. The generation module encapsulates cross-modal interactions to assist in clinical prediction. The **prediction stage** fuses the generated latent CXR, the most recent CXR image, and the complete EHR time series for clinical predictions.

3.1 Notations and Preliminaries

EHR and CXR data   Patient-wisely, we denote the EHR time series within the time interval between t_i and t_j as $\mathbf{X}_{(t_i,t_j)}^{\text{EHR}} = [\mathbf{x}_{t_i}, \mathbf{x}_{t_i+1}, \ldots, \mathbf{x}_{t_j}]$, where $\mathbf{x}_t \in \mathbb{R}^K$ is the variables recorded at time t and K is the number of features. We denote the grayscale CXR images taken at the time t_i as $\mathbf{X}_{t_i}^{\text{CXR}} \in \mathbb{R}^{W \times H}$, where W and H denote its width and height, respectively. For each CXR image, we extract the abnormality finding label, $\mathbf{y}_{t_i}^{\text{CXR}}$, from radiology reports using CheXpert [44].

Predictive latent space for CXR   Using diffusion models in a semantic latent space, rather than a high-dimensional data space, has shown a substantial decrease in computational expenses with minimal impact on synthesis quality [22]. To obtain an informative and expressive latent space, we first train a VAE [45] consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The data used for training VAE are all available CXR images in the training set with corresponding abnormality finding labels, $(\mathbf{X}_t^{\text{CXR}}, \mathbf{y}_t^{\text{CXR}})$. The primary objective of the VAE is to reconstruct the original CXR image $\mathbf{X}_t^{\text{CXR}}$ with $\mathcal{D}(\mathcal{E}(\mathbf{X}_t^{\text{CXR}}))$. We follow the VAE training process in [22], incorporating a pixel-wise reconstruction loss accompanied by a perceptual loss [46], an adversarial objective, and a lightly-penalized Kullback-Leibler loss towards a standard normal aiming at constraining the latent spaces from excessively high variance. Besides, to improve the encoder's ability to predict, we also include a prediction loss regarding the abnormality label $\mathbf{y}_t^{\text{CXR}}$. We denote the encoded latent CXR by $\mathbf{Z}_t = \mathcal{E}(\mathbf{X}_t^{\text{CXR}}) \in \mathbb{R}^{C \times \frac{W}{r} \times \frac{H}{r}}$, where C represents the channel of the compressed representation and r represents the compression ratio. We first pre-train the VAE model and then freeze it throughout the training and inference of DDL-CXR. More details on VAE training are presented in Appendix A.1.

3.2 LDM Stage: Dynamic Latent CXR Generation

As discussed previously, to generate an up-to-date, patient-specific latent CXR, it is important to incorporate the unique anatomical details of the individual patient. Furthermore, the generated image must accurately reflect the evolving pathology as documented in the irregular EHR time series. To this end, we extract all sequential image pairs and the EHR time series between them. We denote each sample as a quadruplet: $(\mathbf{X}_{t_0}^{\text{CXR}}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}, \mathbf{X}_{t_1}^{\text{CXR}}, \mathbf{y}_{t_1}^{\text{CXR}})$. CXR images are encoded using the pre-trained VAE as we aim to generate latent CXR images: $\mathbf{Z}_{t_0} = \mathcal{E}(\mathbf{X}_{t_0}^{\text{CXR}}), \mathbf{Z}_{t_1} = \mathcal{E}(\mathbf{X}_{t_1}^{\text{CXR}})$. We follow prior works on diffusion models to learn our LDM [22, 32]. It comes down

4

to learning a network that predicts the noise added to the noisy latent $\mathbf{Z}_{t_1}^{(n)}$ at denoising step n as $\epsilon_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}), n\right)$. Following prior works [22, 32], we parameterize ϵ_θ by a standard UNet [47]. Here $f_{\text{cond}}^{\text{EHR}}(\cdot)$ is the encoder for the irregular EHR time series to be detailed later. The detailed diffusion and denoising processes are presented in Appendix A.1.

Neural backbone and conditioning mechanisms Due to the remarkable capability of UNet [47] in capturing the spatial structure of images, we follow prior works and use a UNet as our neural backbone ϵ_θ. It predicts the noise added in the diffusion process, conditioned on the reference image and the EHR time series. To explicitly capture and utilize the anatomical structure of individual patients, we first concatenate the reference latent CXR $\mathbf{Z}_{t_0}$ and the step-n noisy latent $\mathbf{Z}_{t_1}^{(n)}$. To further integrate the disease course embedded in the EHR time series, we use the cross-attention mechanism to capture the interaction between the two modalities. Formally, the input to the UNet layers is given by:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V},$$

$$\text{with } \mathbf{Q} = \mathbf{W}_Q \cdot \varphi\left(\mathbf{Z}_{t_1}^{(n)} || \mathbf{Z}_{t_0}\right), \mathbf{K} = \mathbf{W}_K \cdot f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}), \mathbf{V} = \mathbf{W}_V \cdot f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}),$$

(1)

where $\varphi(\cdot)$ denotes the flattened intermediate representation of the UNet and $||$ denotes concatenation.

Capturing disease course via EHR time series To effectively capture useful information on disease progression for future CXR generation, we adopt a multi-task Transformer-based time series encoder [48] with the masked self-attention mechanism to handle the variable length of EHR time series [49]. The encoded representation of EHR, $\mathbf{E}_{(t_0,t_1)}$, is given by

$$\mathbf{E}_{(t_0,t_1)} = f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}) = \text{Transformer}\left([\mathbf{h}_{\text{CLS}}, \phi(\mathbf{x}_{t_0}), \ldots, \phi(\mathbf{x}_{t_1})]\right),$$

(2)

where $\phi(\mathbf{x}_t)$ projects the original EHR time series into an embedding space and applies the positional encoding at time step t. $\mathbf{h}_{\text{CLS}}$ is the class token.

To further extract information that is relevant to CXR generation and facilitate modality fusion at the LDM stage, we incorporate an auxiliary prediction task: using the class token from the encoded EHR to predict the abnormality findings $\mathbf{y}_{t_1}^{\text{CXR}}$, associated with the CXR image $\mathbf{X}_{t_1}^{\text{CXR}}$, i.e., $\widehat{\mathbf{y}}_{t_1}^{\text{CXR}} = g(\mathbf{h}_{\text{CLS}})$, where g denotes the prediction function, e.g., an MLP, which is trained by jointly minimize the loss function given by $\mathcal{L}_{\text{aux}} := \frac{1}{M}\frac{1}{L}\sum_{m=1}^{M}\sum_{l=1}^{L} y_{ml}^{\text{CXR}} \log(\widehat{y}_{ml}^{\text{CXR}}) + (1 - y_{ml}^{\text{CXR}})\log(1 - \widehat{y}_{ml}^{\text{CXR}})$, where M is the number of training samples for LDM and L is the number of classes of abnormality labels of CXR. The auxiliary task enables the EHR encoder to extract CXR-related information, which further encourages the interaction between EHR and CXR to be captured in the subsequent generation.

Enhancing semantic multimodal fusion via contrastive LDM learning The generation conditioning on EHR data is challenging because the EHR and CXR data are highly heterogeneous and the interactions are implicit. To force the LDM to utilize EHR information during generation, we propose a contrastive way of learning the conditional LDM. Specifically, for each EHR time series, we obtain a perturbed version of its representation $\widetilde{\mathbf{E}}_{(t_0,t_1)} = (1 - \beta)\mathbf{E}_{(t_0,t_1)} + \beta\boldsymbol{\delta}$, where $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is randomly drawn from a standard normal distribution, β is a hyperparameter controlling the strength of the noise. When the perturbed EHR is given as input, we expect the generated image to be far away from the target image. This leads to the following training objective function:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathbf{Z}_{t_1}, \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), n}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}), n\right)\right\|_2^2\right.$$

$$\left. + \lambda_1 \max\left(\left\|\epsilon - \epsilon_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, \mathbf{E}_{(t_0,t_1)}, n\right)\right\|_2^2 - \left\|\epsilon - \epsilon_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, \widetilde{\mathbf{E}}_{(t_0,t_1)}, n\right)\right\|_2^2 + \alpha, 0\right)\right],$$

(3)

where α is a hyperparameter controlling the tolerance of the noisy-conditional generation. λ_1 is a coefficient controlling the strength of the contrastive term. To ensure stability during training, we set the initial value of λ_1 to zero and linearly increase it to one during training.

5

3.3 Prediction Stage

In the prediction stage, we do not have access to an up-to-date CXR image. Therefore, we generate an updated latent CXR $\widehat{\mathbf{Z}}_{t_1}$ at the prediction time t_1 using the last available CXR image $\mathbf{X}_{t_0}^{\text{CXR}}$ as the reference image and the EHR time series in between $\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}$. To make predictions using available EHR, we adopt another time series encoder, $f_{\text{pred}}^{\text{EHR}}$, which has the same structure as $f_{\text{cond}}^{\text{EHR}}$ as in Eq. (2). Note that for prediction, the EHR data used, $\mathbf{X}_{\leq 48h}^{\text{EHR}}$ covers all EHR time series with the observation time set as 48 hours, ensuring the available information is fully utilized. In other words, $\mathbf{X}_{(t_0,t_1)}^{\text{EHR}} \subseteq \mathbf{X}_{\leq 48h}^{\text{EHR}}$. In clinical practice, clinicians make predictions not only based on the latest CXR, but also on past CXR images as reference for disease basis. To this end, we employ all available data: $\mathbf{X}_{t_0}^{\text{CXR}}$, $\mathbf{X}_{\leq 48h}^{\text{EHR}}$ and the generated latent CXR $\widehat{\mathbf{Z}}_{t_1}$ to make the final clinical prediction:

$$\widehat{\mathbf{y}} = \mathcal{G}_\psi \left(f_{\text{pred}}^{\text{CXR}}(\mathbf{X}_{t_0}^{\text{CXR}}), \ f_{\text{pred}}^{\text{EHR}}(\mathbf{X}_{\leq 48h}^{\text{EHR}}), \ f_{\text{pred}}^{\text{LAT}}(\widehat{\mathbf{Z}}_{t_1}) \right). \tag{4}$$

Here f_{pred}^{i}, $i \in \{\text{CXR}, \text{EHR}, \text{LAT}\}$ are encoders for CXR, EHR, and the generated latent CXR, accordingly. We parameterize $f_{\text{pred}}^{\text{LAT}}$ and $f_{\text{pred}}^{\text{EHR}}$ using Transformer models, and $f_{\text{pred}}^{\text{CXR}}$ using a ResNet model. The predicting model $\mathcal{G}_\psi$ with ψ denoting the model parameter, is parameterized by a self-attention layer. We learn it by minimizing the cross-entropy (CE) loss:

$$\mathcal{L}_{\text{task}} := \sum_{m=1}^{M'} \sum_{l=1}^{L'} y_{ml} \log(\widehat{y}_{ml}) + (1 - y_{ml}) \log(1 - \widehat{y}_{ml}), \tag{5}$$

where L' is the number of classes in the prediction task and M' is the number of training samples in the prediction stage.

4 Experiments

4.1 Experiment Settings

Datasets   We empirically evaluate the clinical predictive performance of DDL-CXR using MIMIC-IV [50] and MIMIC-CXR [18][4]. MIMIC-IV comprises de-identified critical care data from adult patients admitted to either ICUs or the emergency department (EDs) of Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019, and MIMIC-CXR contains chest X-rays and reports collected from BIDMC, with a subset of patients matched with those in MIMIC-IV. For EHR data, we follow a preprocessing pipeline similar to that described in [10]. 17 clinical time series variables as well as age and gender are extracted. The details can be found in Appendix A.2.

Dataset construction and partition   The inclusion criteria for this study involve ICU stays from the matched subset of MIMIC-IV and MIMIC-CXR that contain at least one CXR image (with Anterior-Posterior (AP) projection) during the ICU stay or within 24 hours before ICU admission. We exclude ICU stays with lengths shorter than 48 hours. The dataset is randomly split by the patient identifier with a ratio of 24:4:7 for training, validation, and testing, which avoids patient overlapping between subsets.

From the training patients, we further extract data for training the VAE, the LDM, and the prediction model. We extract all images from the training patients for training VAE and extract all CXR image pairs of the same patient taken at any interval greater than 12 hours for training the LDM, i.e.,

$$\mathcal{D}_{\text{LDM}} = \left\{ \left(\mathbf{X}_{t_0}^{\text{CXR}}, \mathbf{X}_{t_1}^{\text{CXR}}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}, \mathbf{y}_{t_1}^{\text{CXR}} \right)_{(t_1 - t_0) > 12h} \right\},$$

where a single ICU stay may contain multiple data pairs for LDM training. This greatly enlarges the training subset for the LDM stage.

For the prediction stage, we extract the last available CXR image and the EHR time series in the first 48 hours and the label for the prediction task of each ICU stay, i.e., the triplet $\left(\mathbf{X}_{\text{last}}^{\text{CXR}}, \mathbf{X}_{\leq 48h}^{\text{EHR}}, \mathbf{y}_{\text{task}} \right)$. Note that the EHR time series used in the prediction stage differs from that in the LDM stage in their time interval since they serve for different purposes.

---

[4]Both are open source under the PhysioNet Credentialed Health Data License 1.5.0 license.

6

Table 1: Overall performance for the phenotype classification and mortality prediction task as measured by AUPRC and AUROC scores. `DDL-CXR` outperforms all baselines in these metrics.

	Phenotyping		Mortality	
	AUPRC	AUROC	AUPRC	AUROC
Uni-EHR [23]	0.434 ±0.009	0.720 ±0.006	0.498 ±0.007	0.815 ±0.007
MMTM [52]	0.430 ±0.005	0.715 ±0.003	0.422 ±0.014	0.785 ±0.004
DAFT [9]	0.435 ±0.002	0.720 ±0.003	0.448 ±0.004	0.800 ±0.003
MedFuse [10]	0.437 ±0.001	0.718 ±0.002	0.443 ±0.009	0.793 ±0.003
DrFuse [13]	0.459 ±0.003	0.729 ±0.004	0.460 ±0.004	0.773 ±0.008
GAN-based [53]	0.453 ±0.010	0.728 ±0.008	0.505 ±0.018	0.816 ±0.010
DDL-CXR (ours)	**0.470** ±0.003	**0.740** ±0.002	**0.523** ±0.011	**0.822** ±0.009

We use the same approach to extract the validation subsets for hyperparameter tuning of VAE, LDM, and the prediction model. Note that the testing patients are held out for evaluating prediction performance only, and are not involved in the training and model selection of VAE and LDM.

Prediction tasks and evaluation metrics   We evaluate `DDL-CXR` with two clinical prediction tasks: in-hospital mortality prediction and phenotype classification using clinical data collected within the first 48 hours of ICU admissions. The phenotype classification is a multi-label classification task, where the labels are defined by the 25 disease phenotypes, extracted following [51]. The details of the label prevalence and data cohort statistics can be found in Appendix A.2.

We evaluate the performance using two metrics, the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristics (AUROC). For the phenotyping task, we report macro-averaged scores. We conduct each prediction experiment five times with distinct random seeds and reported the mean and standard deviation of the results.

Baseline Models   We compare the following methods. (1) **Uni-EHR**, a single-modal classifier for EHR time series based on Transformer [23], (2) **MMTM** [52], a multi-modal fusion method based on CNNs through squeeze and excitation operations, (3) **DAFT** [9] a general-purpose module for fusing tabular clinical information and image data by dynamically rescaling and shifting the feature maps of a convolutional layer, (4) **MedFuse** [10], an LSTM-based multimodal fusion method developed for clinical prediction using EHR and CXR, (5) **DrFuse** [13], a disentangled learning approach that handles modality missing and modal inconsistency in clinical multi-modal fusion, and (6) **GAN-based generation** [53], a model originally proposed to generate individual brain images conditioning on age and Alzheimer's Disease (AD) status via training a conditional GAN.

4.2   Prediction Performance

`DDL-CXR` **obtains the best overall performance.**   We summarize the overall performance of the phenotype classification and in-hospital mortality prediction in Table 1, where `DDL-CXR` outperforms all baselines. This shows that generating an updated CXR during test time is beneficial for downstream tasks. On the contrary, DrFuse, MedFuse, DAFT, and MMTM use the last available CXR for prediction, which might have been outdated.

The performance gain of `DDL-CXR` in terms of AUPRC is particularly noteworthy as the AUPRC metric is especially relevant in the context as it underscores the effectiveness of our approach in identifying the positive class in imbalanced medical datasets. `DDL-CXR` achieves relative improvements of 2.4% and 3.56% over the best baselines in terms of AUPRC for phenotype classification and mortality prediction, respectively.

Mortality prediction with varying time interval   We define the time interval (by hour) between the prediction time and the time of the last available CXR as δ and compute the evaluation metrics in patient groups with different ranges of δ. The results are presented in Table 2. Since the label prevalence varies significantly between groups, making the comparison of AUPRC between groups less meaningful, we report AUROC in the paper and AUPRC in the appendix. `DDL-CXR` consistently outperforms the baseline models for most groups of δ for the mortality prediction task. As the δ increases, the last CXR becomes more "outdated", and we observe a noticeable increase in the

7

Table 2: The mean of AUROC score with standard deviation for mortality prediction for overall and different time gaps. δ represents the time interval (by hour) between the prediction time and the time of the last available CXR. Numbers in bold indicate the best performance in each column. DDL-CXR outperforms all baselines in most settings. The AUPRC scores can be found in Appendix B.1.

	Overall	$\delta < 12$	$12 \leq \delta < 24$	$24 \leq \delta < 36$	$\delta \geq 36$
prevalence	14.7%	16.6%	19%	15.9%	9.26%
Uni-EHR [23]	0.815 ±0.007	0.854 ±0.010	0.799 ±0.013	0.756 ±0.019	0.796 ±0.008
MMTM [52]	0.785 ±0.004	0.798 ±0.008	0.763 ±0.004	0.760 ±0.012	0.772 ±0.014
DAFT [9]	0.800 ±0.003	0.803 ±0.010	0.782 ±0.009	**0.776** ±0.006	0.796 ±0.008
MedFuse [10]	0.793 ±0.003	0.812 ±0.004	0.762 ±0.007	0.760 ±0.009	0.800 ±0.010
DrFuse [13]	0.773 ±0.008	0.802 ±0.012	0.717 ±0.023	0.757 ±0.041	0.723 ±0.013
GAN-based [53]	0.816 ±0.010	0.846 ±0.010	**0.800** ±0.011	0.760 ±0.026	0.806 ±0.016
DDL-CXR (ours)	**0.822** ±0.009	**0.867** ±0.015	**0.800** ±0.008	0.753 ±0.015	**0.830** ±0.011

Table 3: The AUPRC score of predicting each phenotype label. DDL-CXR obtains the highest average rank. Full names of phenotype labels and AUROC scores can be found in the Appendix.

	Uni-EHR	MMTM	DAFT	MedFuse	DrFuse	GAN-based	DDL-CXR
Acute renal failure	*0.573*	0.568	0.572	0.565	0.564	0.563	**0.588**
Acute cerebrovascular disease	0.425	0.418	0.419	*0.434*	0.399	**0.446**	0.416
Acute myocardial infarction	0.185	0.192	0.187	**0.219**	*0.209*	0.171	0.206
Cardiac dysrhythmias	0.579	0.532	0.548	0.560	*0.584*	0.561	**0.605**
Chronic kidney disease	0.515	0.505	*0.515*	0.497	0.477	0.501	**0.538**
COPD and bronchiectasis	0.319	0.327	0.342	0.344	**0.405**	0.372	*0.382*
Surgical complications	0.370	0.379	*0.385*	0.381	0.377	0.344	**0.388**
Conduction disorders	0.276	0.287	0.298	0.286	*0.632*	0.609	**0.633**
CHF; nonhypertensive	0.593	0.619	0.647	0.631	*0.661*	0.652	**0.682**
CAD	0.560	0.540	0.556	0.544	0.581	*0.590*	**0.611**
DM with complications	*0.562*	**0.569**	0.552	0.561	0.550	0.552	0.524
DM without complication	**0.370**	0.367	0.343	0.356	*0.369*	0.352	0.368
Disorders of lipid metabolism	*0.594*	0.576	0.570	0.566	0.584	0.587	**0.601**
Essential hypertension	0.551	0.519	0.525	0.518	0.502	*0.554*	**0.561**
Fluid and electrolyte disorders	0.655	*0.664*	0.662	0.656	0.658	0.662	**0.672**
Gastrointestinal hemorrhage	0.180	0.142	0.162	**0.192**	*0.191*	0.151	0.180
Secondary hypertension	*0.463*	0.455	0.452	0.453	0.437	0.451	**0.484**
Other liver diseases	0.316	0.316	0.341	0.344	*0.372*	0.362	**0.378**
Other lower respiratory disease	0.219	0.209	0.206	0.223	**0.255**	0.236	*0.242*
Other upper respiratory disease	0.166	0.137	0.166	0.202	**0.274**	0.196	*0.234*
Pleurisy; pneumothorax	0.143	0.145	0.159	0.159	**0.172**	*0.171*	0.166
Pneumonia	0.412	**0.437**	*0.429*	0.419	0.406	0.415	0.428
Respiratory failure	0.655	*0.686*	0.674	0.671	**0.692**	0.663	0.669
Septicemia (except in labor)	*0.585*	0.573	0.580	0.565	0.562	0.573	**0.603**
Shock	*0.590*	0.584	0.582	**0.592**	0.572	0.587	0.586
Average Rank	4.4	4.64	4.4	4.24	*3.88*	4.16	**2.28**

performance gap between the best baseline and DDL-CXR in the group ($\delta \geq 36$), from 0.806 to 0.830. This validates our hypothesis that the generation of a timely CXR, accounting for disease progression, can significantly enhance the performance of clinical predictions.

Phenotype classification The class-wise AUPRC scores for the phenotyping task are detailed in Table 3, where DDL-CXR demonstrates notable performance improvements, achieving the highest average rank across all phenotype labels. Due to space limit, we report the standard deviations and the AUROC scores in Appendix B.2. The improvement over baseline multimodal fusion methods validates the effectiveness of facilitating fusion between EHR and CXR in the presence of asynchronicity.

8

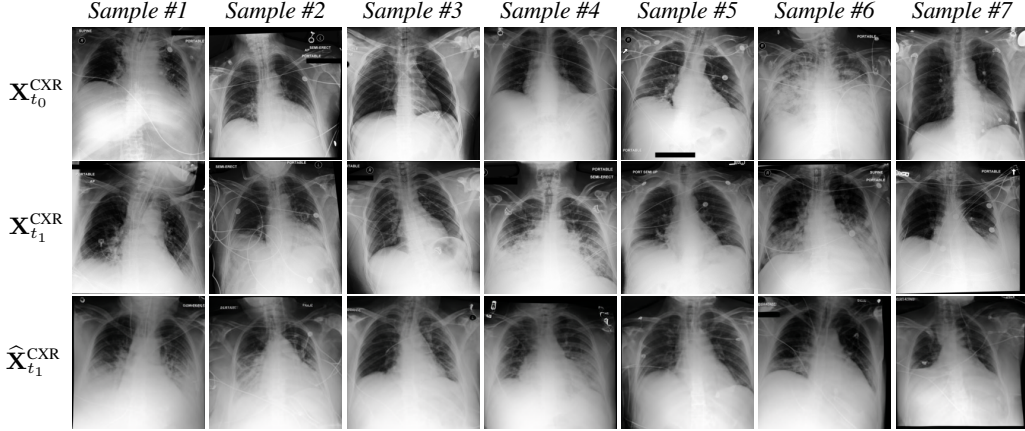



Figure 3: Examples of images generated by DDL-CXR. From top to bottom, the three rows are reference images $\mathbf{X}_{t_0}^{\text{CXR}}$, ground-truth images $\mathbf{X}_{t_1}^{\text{CXR}}$, and generated images $\widehat{\mathbf{X}}_{t_1}^{\text{CXR}}$, respectively. The generations show that DDL-CXR captures the anatomical information from $\mathbf{X}_{t_0}^{\text{CXR}}$ and the information of disease progression extracted from EHR is blended well towards generating $\mathbf{X}_{t_1}^{\text{CXR}}$.

4.3 Quality of Generated Chest X-ray Images

Quantitative Evaluation   We evaluate the quality of generated CXRs using the test set of the LDM stage, where the ground-truth target CXR is available. The Fréchet Inception Distance (FID) score [54] evaluates the similarity between the distributions of generated and ground-truth target CXRs by computing Fréchet distance on the representation obtained from a pre-trained Inception-v3 network.

Besides, we directly measure the Fréchet distance (FD) and Wasserstein distance (WD) in the latent space of the VAE between the generated and the target CXRs. Results are shown in Table 4. Results of "Last-CXR" are obtained between reference images $\mathbf{X}_0$ and target images $\mathbf{X}_1$, both are directly from the dataset without generation. Thus, this provides a reference to the lower bound of the metrics used. DDL-CXR surpasses GAN-based methods across all metrics and obtains the lowest FD and WD. "w/o $\mathbf{Z}_{t_0}$" and "w/o $\mathbf{E}_{(t_0,t_1)}$" are obtained by removing the condition of the last available CXR and EHR data, respectively. Notably,

Table 4: Generation quality.

	FID ($\downarrow$)	FD ($\downarrow$)	WD ($\downarrow$)
Last-CXR	16.50	2322.68	6353.06
w/o $\mathbf{Z}_{t_0}$	47.91	3260.17	7491.17
w/o $\mathbf{E}_{(t_0,t_1)}$	**30.03**	2412.80	7226.38
GAN-based	98.67	3651.27	7922.71
DDL-CXR	33.83	**2316.08**	**7132.82**

excluding EHR data from the generation conditions resulted in lower FID scores, which is natural since the generation becomes less restrictive.

Qualitative Evaluation   To further visually examine the generated CXR images, we decode the latent CXR and visualize seven examples in Fig. 3. The first row shows the last CXR images used as reference, the second row displays the ground-truth CXR images, and the last row showcases the generated CXR images. The comparison between the first and third rows indicates that the generated CXR could well capture anatomical structure, while the comparison between the second and third rows demonstrates that the generated CXRs are in line with the latest imaging manifestations, implying that the disease progression embedded in EHR could be captured and utilized in the generation process. We further retrieve the radiology reports and the discharge summary of the corresponding patients from the database for case studies. Fig. 1 shows one example of the case study (*Sample #4*), where the patient rapidly turned from normal CXR to severe pulmonary consolidation. The discharge summary shows that the patient experienced transfusion-related acute lung injury and sepsis. Evidently, the generated CXR could more accurately reflect the progressed condition of the patient. Due to space limits, we present more case studies in Appendix B.5.

4.4 Ablation Study

To better understand the factors contributing to the improved performance, we conducted an ablation study by removing the conditioning components in the LDM stage. The results are summarized

9

Table 5: Results of the ablation study.

	Phenotyping		Mortality	
	AUPRC	AUROC	AUPRC	AUROC
Last-CXR	0.459 ±0.012	0.730 ±0.008	0.503 ±0.010	0.817 ±0.007
w/o $\mathbf{Z}_{t_0}$	0.448 ±0.012	0.726 ±0.008	0.494 ±0.014	0.811 ±0.008
w/o $\mathbf{E}_{(t_0,t_1)}$	0.461 ±0.002	0.723 ±0.006	0.474 ±0.016	0.799 ±0.015
w/o Contrastive	0.460 ±0.007	0.722 ±0.011	0.483 ±0.019	0.802 ±0.011
w/o $\mathcal{L}_{\text{aux}}$	0.461 ±0.003	0.718 ±0.012	0.495 ±0.026	0.811 ±0.009
Last-CXR (w/o EHR)	0.376 ±0.009	0.661 ±0.007	0.243 ±0.002	0.664 ±0.006
DDL-CXR (w/o EHR)	0.385 ±0.006	0.668 ±0.007	0.269 ±0.013	0.707 ±0.008
DDL-CXR	**0.470** ±0.003	**0.740** ±0.002	**0.523** ±0.011	**0.822** ±0.009

in Table 5. The variant "Last-CXR" has the same architecture as the classifier of DDL-CXR but removes the generated latent CXR $\widehat{\mathbf{Z}}_{t_1}$. The improvement over Last-CXR shows that learning an LDM for generating updated latent CXR is a more effective approach to multimodal fusion, and hence benefits downstream prediction, especially for the mortality prediction task. The variants "w/o $\mathbf{Z}_{t_0}$" and "w/o $\mathbf{E}_{(t_0,t_1)}$" remove the last available CXR and EHR data, respectively, from the condition during LDM training. "w/o Contrastive" removes the contrastive terms from the LDM objective function. "w/o $\mathcal{L}_{\text{aux}}$" removes the auxiliary loss which drives the EHR encoder to capture the CXR-related abnormality findings. The results show that adding each component brings slight improvement while incorporating the reference CXR, the EHR, the contrastive learning, and the auxiliary task achieves the best performance. We also remove the EHR data completely in the prediction stage and evaluate the performance using the last available CXR and the generated CXR, respectively. Results are shown as "Last-CXR (w/o EHR)" and "DDL-CXR (w/o EHR)" in Table 5. The results suggest that the generation of an updated CXR significantly benefits downstream clinical predictions. Additional experiment results on robustness against reduced training data size can be found in Appendix B.4.

5 Broader Impacts and Limitations

DDL-CXR holds promise for societal benefits, such as more precise and timely medical interventions, and offers an alternative for patients with limited access to X-ray imaging. Nonetheless, the potential for generating fake profiles necessitates stringent safeguards, including expert validation of synthesized images, to prevent misuse and protect patient confidentiality, especially when applied to private datasets. Despite its promise, DDL-CXR has some limitations like the need for meticulous hyperparameter tuning and a performance gap across different time intervals, as indicated in Table 2. Addressing such potential biases is a priority for future research. Furthermore, while various metrics have been employed to assess generation quality, expert evaluation by radiologists would provide a more insightful measure of the model's efficacy.

6 Conclusion

In this paper, we introduce DDL-CXR, which utilizes a powerful LDM to dynamically generate up-to-date latent chest X-rays to tackle the asynchronicity of multi-modal clinical data for predictions. Our approach involves leveraging various conditions for patient-specific generation: the most recent available chest X-ray to incorporate detailed patient-specific anatomical structure, as well as the EHR data with variable durations for disease progression information. To improve multi-modal fusion in the generation, we develop a contrastive-learning-based LDM to capture and utilize disease courses in EHR. Through quantitative and qualitative validations, we demonstrate the superior performance of DDL-CXR in both image generation and enhancing multi-modal fusion via conditional generation for clinical prediction.

10

Acknowledgments and Disclosure of Funding

This work is partially supported by the General Research Fund of Hong Kong Research Grants Council (project no. 15218521), a grant under Theme-based Research Scheme of Hong Kong Research Grants Council (project no. T45-401/22-N), the General Research Fund RGC/HKBU12202621 from the Research Grant Council, the Research Matching Grant Scheme RMGS2021_8_06 from the Hong Kong Government, the National Natural Science Foundation of China (62302413), and the Health and Medical Research Fund (23220312).

References

[1] Rowa Aljondi and Salem Alghamdi. Diagnostic value of imaging modalities for COVID-19: scoping review. *Journal of Medical Internet Research*, 22(8):e19673, 2020.

[2] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981, 2022.

[3] Kejing Yin, Dong Qian, and William K Cheung. PATNet: Propensity-adjusted temporal network for joint imputation and prediction using binary EHRs with observation bias. *IEEE Transactions on Knowledge & Data Engineering*, 36(06):2600–2613, 2024.

[4] Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. Learning inter-modal correspondence and phenotypes from multi-modal electronic health records. *IEEE Transactions on Knowledge & Data Engineering*, 34(09):4328–4341, 2022.

[5] Kejing Yin, Ardavan Afshar, Joyce C Ho, William K Cheung, Chao Zhang, and Jimeng Sun. LogPar: Logistic PARAFAC2 factorization for temporal binary data with missing values. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1625–1635, 2020.

[6] Lihong Song, Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. Medical concept embedding with multiple ontological representations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4613–4619, 2019.

[7] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 3(1):136, 2020.

[8] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P Lungren. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1):22147, 2020.

[9] 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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 688–698. Springer, 2021.

[10] 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, 05–06 Aug 2022.

[11] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[12] Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, 2023.

[13] 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.

[14] 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.

11

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

[16] Claudia I Henschke, David F Yankelevitz, Austin Wand, Sheila D Davis, and Maria Shiau. Accuracy and efficacy of chest radiography in the intensive care unit. *Radiologic Clinics of North America*, 34(1):21–31, 1996.

[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] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[19] Anusoumya Ganapathy, Neill KJ Adhikari, Jamie Spiegelman, and Damon C Scales. Routine chest X-rays in intensive care units: a systematic review and meta-analysis. *Critical Care*, 16(2):1–12, 2012.

[20] Declan Grant, Bartłomiej W Papież, Guy Parsons, Lionel Tarassenko, and Adam Mahdi. Deep learning classification of cardiomegaly using combined imaging and non-imaging ICU data. In *Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25*, pages 547–558. Springer, 2021.

[21] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 23–29 Jul 2023.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Chaoqi Yang, M Brandon Westover, and Jimeng Sun. ManyDG: Many-domain generalization for healthcare applications. In *The Eleventh International Conference on Learning Representations*, 2023.

[25] Zhicheng Jiao, Ji Whae Choi, Kasey Halsey, Thi My Linh Tran, Ben Hsieh, Dongcui Wang, Feyisope Eweje, Robin Wang, Ken Chang, Jing Wu, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest X-rays and clinical data: a retrospective study. *The Lancet Digital Health*, 3(5): e286–e294, 2021.

[26] Zhuo Zhi, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. Multimodal diagnosis for pulmonary embolism from EHR data and CT images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2053–2057. IEEE, 2022.

[27] Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Margot Brereton, Jacinto C Nascimento, Joaquim Jorge, and Catarina Moreira. MDF-Net for abnormality detection by fusing X-rays with clinical data. *Scientific Reports*, 13(1):15873, 2023.

[28] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett Landman, and Yuankai Huo. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 2023.

[29] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical Tansformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer, 2022.

[30] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

12

[31] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027, June 2023.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[35] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[36] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023.

[37] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022. *URL https://arxiv. org/abs/2204.06125*, 7, 2022.

[39] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[40] Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M Pohl. Generating realistic brain MRIs via a conditional diffusion probabilistic model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2023.

[41] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[42] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest X-ray synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 180–191. Springer, 2023.

[43] Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

[44] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[45] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[46] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

13

[48] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A Tansformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[50] 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.

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

[52] 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.

[53] Tian Xia, Agisilaos Chartsias, Chengjia Wang, Sotirios A Tsaftaris, Alzheimer's Disease Neuroimaging Initiative, et al. Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, 73:102169, 2021.

[54] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[55] Mohammad Amin Morid, Olivia R Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, 2023.

14

A Experiment Details

A.1 Details of Architectures and Training Procedures of `DDL-CXR`

The training and validation processes are executed on a server equipped with a RTX 4090-24GB GPU card and a 16 vCPU Intel Xeon Processor. The method is implemented using PyTorch 1.9.1 and PyTorch-Lightning 1.4.2. DDIM [33] sampling with 200 steps is employed to accelerate the sampling process, and AdamW optimizer is used for all model training. Our implementation is partially based on the repository of the latent diffusion model [22][5].

Variational autoencoder (VAE)   The VAE training process, as outlined in [22], includes a pixel-wise reconstruction loss, a perceptual loss [46], an adversarial objective, and a lightly-penalized Kullback-Leibler loss toward a standard normal to constrain latent spaces, given by:

$$\mathcal{L}_{\text{VAE}} = \min_{\mathcal{E}, \mathcal{D}, \Phi} \max_{\psi} \Big(L_{\text{rec}}(\mathbf{X}^{\text{CXR}}, \mathcal{D}(\mathcal{E}(\mathbf{X}^{\text{CXR}}))) - L_{\text{adv}}(\mathcal{D}(\mathcal{E}(\mathbf{X}^{\text{CXR}}))) + \log D_{\omega}(\mathbf{X}^{\text{CXR}}))$$
$$+ L_{\text{reg}}(\mathbf{X}^{\text{CXR}}); \mathcal{E}, \mathcal{D}) + L_{\text{CE}}(\Phi(\mathcal{E}(\mathbf{X}^{\text{CXR}})), \mathbf{y}) \Big), \tag{6}$$

where Φ is a classifier that predicts the CheXpert labels associated with the image and we parameterize it with an MLP. All CXRs in the training subset of the LDM stage are gathered for VAE training. A compression rate $r = 8$ is adopted, and the training continues for a maximum of 50 epochs. The model with the minimum validation error, as measured using CXRs from the validation subset, is selected. The resulting latent representation has a dimension of $4 \times 28 \times 28 = 3136$. To restrict the normal prior in the latent space and prioritize reconstruction quality, a KL-divergence weighting of 1×10^{-6} is set. We use the base learning rate of 4.5×10^{-6}, which is scaled by the number of GPU cards and batch size.

Latent diffusion model (LDM) stage in `DDL-CXR`   In the LDM stage of our `DDL-CXR` model, we employ the UNet architecture [47] as the neural backbone, denoted by ϵ_{θ}. Meanwhile, we utilize a multivariate time series Transformer [48] for the EHR conditioning encoder $f_{\text{cond}}^{\text{EHR}}$. The Transformer $f_{\text{cond}}^{\text{EHR}}$ is designed with one layer, a model dimension d set to 128, and a maximum EHR data length of 70. The UNet model ϵ_{θ} features an input channel of 8 and an output channel of 4. The encoding section comprises three blocks, with model channels set at 224, 448, and 672, consisting of a ResBlock module followed by a spatial transformer. The bottleneck consists of two ResBlock modules with a spatial transformer in between. The decoder mirrors the encoder architecture. As discussed in Section 3.2, we introduce the encoded EHR information through multi-head cross-attention to the spatial transformer module of ϵ_{θ}. The context dimension is set to 128, and the number of attention heads is 8. The model is trained for 200 epochs with a batch size of 32, and the model with the smallest composite loss on the validation set is selected for subsequent latent Chest X-ray (CXR) generation. We set the hyperparameters α to 0.2, and β to 0.5, empirically.

Latent CXR generation via LDM   In the LDM stage, we aim to generate latent CXR images at time t_1, conditioned on $\mathbf{X}_{t_0}^{\text{CXR}}$ and $\mathbf{X}_{(t_0, t_1)}^{\text{EHR}}$. We first encode the CXR images using the pre-trained VAE, given by:

$$\mathbf{Z}_{t_0} = \mathcal{E}(\mathbf{X}_{t_0}^{\text{CXR}}), \quad \mathbf{Z}_{t_1} = \mathcal{E}(\mathbf{X}_{t_1}^{\text{CXR}}). \tag{7}$$

Essentially, the latent CXR generation requires us to estimate the underlying data distribution $q(\mathbf{Z}_{t_1} | \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0, t_1)}^{\text{EHR}})$. The LDM approximates this distribution via a model distribution $p_{\theta}(\mathbf{Z}_{t_1}^{(0)} | \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0, t_1)}^{\text{EHR}})$, where $\mathbf{Z}_{t_1}^{(0)}$ represents the prior of a CXR in the latent space encoded by the VAE.

We follow prior work on diffusion models to learn our LDM, which involves two processes [32, 22]. In the diffusion process, we gradually add Gaussian noise to $\mathbf{Z}_{t_1}^{(0)}$ in N steps, producing a sequence of noisy representations $\mathbf{Z}_{t_1}^{(1)}, \mathbf{Z}_{t_1}^{(2)}, ..., \mathbf{Z}_{t_1}^{(N)}$, with the transition probability given by:

$$q(\mathbf{Z}_{t_1}^{(n)} | \mathbf{Z}_{t_1}^{(n-1)}) = \mathcal{N}(\mathbf{Z}_{t_1}^{(n)}; \sqrt{1 - \beta_n} \mathbf{Z}_{t_1}^{n-1}, \beta_n \mathbf{I}),$$
$$q(\mathbf{Z}_{t_1}^{(n)} | \mathbf{Z}_{t_1}^{(0)}) = \mathcal{N}(\mathbf{Z}_{t_1}^{(n)}; \sqrt{\bar{\alpha}_n} \mathbf{Z}_{t_1}^{(0)}, (1 - \bar{\alpha}_n)\boldsymbol{\epsilon}), \tag{8}$$

---

[5]`https://github.com/CompVis/latent-diffusion`, open source under MIT license.

15

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents the added noise. The noise level is represented by $\bar{\alpha}_n := \prod_{s=1}^{n}(1-\beta_s)$, where $\{\beta_n \in (0,1)\}_{n=1}^{N}$ is a pre-defined variance schedule. At step N, $\mathbf{Z}_{t_1}^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ follows an isotropic Gaussian distribution. In the denoising process, we start with a sample from the isotropic Gaussian distribution $\mathbf{Z}_{t_1}^{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and gradually recreate the latent CXR $\mathbf{Z}_{t_1}^{(0)}$, conditioned on the reference latent CXR, $\mathbf{Z}_{t_0}$, and the EHR data in between, $\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}$. The generation process is given by:

$$p_\theta\left(\mathbf{Z}_{t_1}^{(0:N)}|\mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}\right) = p\left(\mathbf{Z}_{t_1}^{(N)}\right) \prod_{n=1}^{N} p_\theta\left(\mathbf{Z}_{t_1}^{(n-1)}|\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}\right), \tag{9}$$

where

$$p_\theta\left(\mathbf{Z}_{t_1}^{(n-1)}|\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}\right) = \mathcal{N}\left(\mathbf{Z}_{t_1}^{(n-1)}; \boldsymbol{\mu}_\theta\left(\mathbf{Z}_{t_1}^{(n)}, n, \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}\right), \boldsymbol{\sigma}_n^2 \mathbf{I}\right), \tag{10}$$

and the mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\sigma}_n^2$ are parameterized by

$$\boldsymbol{\mu}_\theta\left(\mathbf{Z}_{t_1}^{(n)}, n, \mathbf{Z}_{t_0}, \mathbf{X}_{(t_0,t_1)}^{\text{EHR}}\right) = \frac{1}{\sqrt{\alpha_n}}\left(\mathbf{Z}_{t_1}^{(n)} - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\boldsymbol{\epsilon}_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}), n\right)\right), \tag{11}$$

and

$$\boldsymbol{\sigma}_n^2 = \frac{1-\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n}\beta_n, \tag{12}$$

where $f_{\text{cond}}^{\text{EHR}}(\cdot)$ is the encoder for the irregular EHR time series. $\boldsymbol{\epsilon}_\theta\left(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_0}, f_{\text{cond}}^{\text{EHR}}(\mathbf{X}_{(t_0,t_1)}^{\text{EHR}}), n\right)$ denotes the generation noise predicted by the neural backbone.

Prediction stage in `DDL-CXR` In the prediction stage, the EHR data is encoded using a one-layer Transformer with a model dimension of 128. We set the dimension of the feedforward layers to 512. The context dimension is also set to 128, and the number of attention heads is 8. We use another Transformer with the same architecture to encode the generated latent CXR $\mathbf{Z}_1$. We use a ResNet-34 model to encode the last available CXR image $\mathbf{X}_0$. The encoded EHR, the encoded latent CXR $\mathbf{Z}_1$, as well as the encoded $\mathbf{X}_0$ are fed into a self-attention layer for final prediction.

A.2 Details of Data Preprocessing

EHR data preprocess We follow a similar EHR data extraction and processing pipeline as [10] but change the sampling frequency from 2h to 1h and introduce two static variables, age, and gender. We extract 17 clinical time series variables (12 continuous and 5 categorical) with discretization and standardization processes exactly the same as in [10]. In addition to the 17 clinical time series variables mentioned in the paper [10], e.g. five categorical (capillary refill rate, Glasgow Coma Scale eye-opening, Glasgow Coma Scale motor response, Glasgow Coma Scale verbal response, and Glasgow Coma Scale total) and 12 continuous (diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH), we introduce two crucial static variables (age and gender) to represent the demographic information of a patient, as patient demographic information is vital for achieving accurate predictions [55]. To construct our dataset, we sampled time series data at hourly intervals, followed by discretization and standardization processes. We adopt masks to handle missing values in time series to capture the missing pattern, acknowledging that the absence of medical data might be intentional and non-random, driven by caregivers [55].

Data Cohort and Potential Selection Bias We summarize the number of samples in the LDM stage and the prediction stage in Table 6. The label prevalences of the two prediction tasks are summarized in Tables 7 and 8. We include the disease prevalence in Table 8 and data cohort statistics in Table 9. Here we summarize a few key points:

- The statistics of the clinical variables are close to each other, suggesting that patients in our data cohort generally have a similar distribution as that in the entire database.
- The overall disease phenotype prevalence is similar.
- For a few thorax-related diseases, our data cohort has a slightly higher prevalence, suggesting that potential selection bias exists.

16

Table 6: Data statistics in training, validation, and testing sets for each stage.

Stage	Training	Validation	Test
LDM stage	8545	1392	2353
Prediction stage	5142	861	1483

Table 7: Label prevalence of in-hospital mortality prediction task.

	Training	Validation	Test
Negative	4396	737	1264
Positive	746	124	219
Negative/Positive	5.89	5.94	5.77

Table 8: Number of samples and prevalence of disease phenotypes in training, validation, and testing sets during the prediction stage. The prevalence of disease phenotypes among all ICU stays from MIMIC-IV database having LoS $\geq$ 48h is given in the last column.

Disease Label	Training	Validation	Testing	Training Prevalence	Validation Prevalence	Testing Prevalence	MIMIC-IV Prevalence
Acute and unspecified renal failure	1932	342	561	0.38	0.40	0.38	0.34
Acute cerebrovascular disease	467	88	113	0.09	0.10	0.08	0.07
Acute myocardial infarction	449	79	131	0.09	0.09	0.09	0.09
Cardiac dysrhythmias	2049	361	601	0.40	0.42	0.41	0.38
Chronic kidney disease	1258	216	399	0.24	0.25	0.27	0.23
Chronic obstructive pulmonary disease	860	151	261	0.17	0.18	0.18	0.16
Complications of surgical/medical care	1218	209	372	0.24	0.24	0.25	0.25
Conduction disorders	581	92	176	0.11	0.11	0.12	0.12
Congestive heart failure; nonhypertensive	1674	288	489	0.33	0.33	0.33	0.30
Coronary atherosclerosis and related	1605	257	491	0.31	0.30	0.33	0.33
Diabetes mellitus with complications	646	105	204	0.13	0.12	0.14	0.12
Diabetes mellitus without complication	1068	182	325	0.21	0.21	0.22	0.18
Disorders of lipid metabolism	2047	349	595	0.40	0.41	0.40	0.41
Essential hypertension	2255	383	611	0.44	0.44	0.41	0.42
Fluid and electrolyte disorders	2685	455	749	0.52	0.53	0.51	0.45
Gastrointestinal hemorrhage	354	68	129	0.07	0.08	0.09	0.07
Hypertension with complications	1129	202	361	0.22	0.23	0.24	0.24
Other liver diseases	899	147	279	0.17	0.17	0.19	0.15
Other lower respiratory disease	718	109	212	0.14	0.13	0.14	0.12
Other upper respiratory disease	376	67	96	0.07	0.08	0.06	0.07
Pleurisy; pneumothorax; pulmonary collapse	553	99	141	0.11	0.11	0.10	0.09
Pneumonia	1149	187	333	0.22	0.22	0.22	0.18
Respiratory failure; insufficiency; arrest (adult)	1741	307	506	0.34	0.36	0.34	0.25
Septicemia (except in labor)	1386	224	425	0.27	0.26	0.29	0.21
Shock	1157	192	357	0.23	0.22	0.24	0.18

Table 9: Statistics of data cohort. We report mean±std for continuous variables and the mode for categorical variables.

Variables	DDL-CXR Cohort	MIMIC-IV Database
Sample size	5142	26330
Age	63.9±16.6	64.3±16.4
Gender (Men, %)	54.98	55.99
Diastolic blood pressure	62.2±10.1	62.8±10.5
Fraction inspired oxygen	0.4±0.2	0.4±0.2
Glucose	141.1±41.7	139±39.6
Heart Rate	87.3±15.4	85.6±15
Height	170±1.1	170±1.5
Mean blood pressure	77.1±10	78.3±10.4
Oxygen saturation	96.7±2	96.5±2
Respiratory rate	19.8±3.7	19.5±3.7
Systolic blood pressure	118.3±15.2	118.6±15.6
Temperature	37±0.4	36.9±0.4
Weight	80.8±20	81.5±19.5
pH	7.2±0.3	7.2±0.3
Capillary refill rate	Normal	Normal
GCS - eye opening	Spontaneously	Spontaneously
GCS - motor response	Obeys Commands	Obeys Commands
GCS – verbal response	Oriented	Oriented
GCS - total	15	15
In-hospital mortality (%)	15	12

17

B Additional Results

B.1 AUPRC of Mortality Prediction

We summarize the AUPRC score for the mortality prediction task in Table 10.

Table 10: The AUPRC score for mortality prediction for overall and different time gaps. δ represents the time gap between the generation (or prediction) time and the occurrence time of the last available CXR. Numbers in bold indicate the best performance in each column. DDL-CXR outperforms all baselines in overall and $\delta < 24$h settings.

	Overall	$\delta < 12$	$12 \leq \delta < 24$	$24 \leq \delta < 36$	$\delta \geq 36$
prevalence	14.7%	16.6%	19%	15.9%	9.26%
Uni-EHR [23]	0.498 ±0.007	0.579 ±0.009	0.541 ±0.016	0.416 ±0.018	0.411 ±0.005
MMTM [52]	0.422 ±0.014	0.430 ±0.010	0.458 ±0.020	0.476 ±0.021	0.374 ±0.012
DAFT [9]	0.448 ±0.004	0.460 ±0.011	0.478 ±0.010	**0.508** ±0.015	0.395 ±0.007
MedFuse [10]	0.443 ±0.009	0.498 ±0.013	0.484 ±0.023	0.432 ±0.015	0.355 ±0.008
DrFuse [13]	0.460 ±0.004	0.506 ±0.033	0.356 ±0.034	0.382 ±0.072	0.399 ±0.019
GAN-based [53]	0.505 ±0.018	0.567 ±0.015	0.551 ±0.018	0.414 ±0.041	**0.442** ±0.034
DDL-CXR (ours)	**0.523** ±0.011	**0.610** ±0.015	**0.566** ±0.013	0.429 ±0.009	0.440 ±0.008

B.2 AUROC of Phenotype Prediction by Disease Label

We summarize the AUROC score for the phenotype prediction task by disease label in Table 11.

Table 11: The AUROC score by disease labels. Results show that DDL-CXR effectively tackles the asynchronicity issue in clinical multi-modal fusion, achieving the highest average rank across all disease labels.

	Uni-EHR [23]	MMTM [52]	DAFT [9]	MedFuse [10]	DrFuse [13]	GAN-based [53]	DDL-CXR (ours)
Acute renal failure	*0.718*±0.003	0.706±0.004	0.712±0.003	0.706±0.004	0.702±0.003	0.710±0.005	**0.723**±0.004
Acute cerebrovascular disease	0.880±0.004	*0.893*±0.003	0.888±0.009	**0.895**±0.005	0.873±0.009	0.875±0.011	0.870±0.008
Acute myocardial infarction	0.712±0.006	0.717±0.005	0.722±0.009	0.724±0.009	*0.730*±0.013	0.701±0.004	**0.735**±0.009
Cardiac dysrhythmias	0.677±0.005	0.648±0.004	0.656±0.004	0.667±0.010	*0.682*±0.006	0.673±0.005	**0.703**±0.007
Chronic kidney disease	0.745±0.008	0.735±0.008	0.732±0.004	0.720±0.004	0.710±0.011	*0.748*±0.005	**0.768**±0.006
COPD and bronchiectasis	0.706±0.006	0.689±0.014	0.707±0.012	0.686±0.006	0.733±0.011	*0.738*±0.003	**0.740**±0.003
Surgical complications	0.636±0.007	*0.656*±0.008	0.648±0.006	**0.662**±0.008	0.649±0.004	0.610±0.013	0.649±0.003
Conduction disorders	0.741±0.006	0.717±0.011	0.735±0.012	0.724±0.008	*0.847*±0.011	0.842±0.006	**0.860**±0.005
CHF; nonhypertensive	0.771±0.004	0.777±0.010	0.796±0.007	0.779±0.008	0.797±0.004	*0.798*±0.006	**0.820**±0.005
Coronary atherosclerosis & others	0.740±0.005	0.728±0.003	0.733±0.007	0.721±0.008	0.742±0.006	*0.753*±0.007	**0.764**±0.006
DM with complications	**0.858**±0.006	0.853±0.006	0.851±0.002	0.845±0.005	0.849±0.006	0.851±0.005	*0.855*±0.004
DM without complication	*0.716*±0.005	0.707±0.007	0.694±0.015	0.704±0.003	0.712±0.008	0.707±0.008	**0.717**±0.009
Disorders of lipid metabolism	0.704±0.007	0.677±0.005	0.676±0.006	0.677±0.003	0.696±0.008	*0.708*±0.006	**0.714**±0.007
Essential hypertension	*0.662*±0.007	0.609±0.002	0.625±0.005	0.632±0.006	0.612±0.003	0.659±0.008	**0.667**±0.008
Fluid and electrolyte disorders	0.671±0.002	0.674±0.003	0.671±0.003	0.665±0.006	0.671±0.004	*0.674*±0.007	**0.677**±0.007
Gastrointestinal hemorrhage	0.664±0.005	0.635±0.017	0.653±0.018	0.664±0.005	*0.677*±0.013	0.662±0.008	**0.678**±0.009
Secondary hypertension	0.725±0.007	0.726±0.007	0.722±0.005	0.716±0.004	0.714±0.010	*0.734*±0.005	**0.752**±0.004
Other liver diseases	0.681±0.007	0.680±0.005	0.693±0.006	0.690±0.004	0.697±0.009	*0.700*±0.010	**0.708**±0.007
Other lower respiratory disease	0.596±0.004	0.606±0.007	0.608±0.003	0.615±0.006	**0.634**±0.015	*0.619*±0.019	0.614±0.011
Other upper respiratory disease	0.691±0.009	0.691±0.009	0.711±0.006	*0.726*±0.010	**0.741**±0.013	0.699±0.019	0.700±0.017
Pleurisy; pneumothorax	0.634±0.005	0.647±0.013	*0.674*±0.014	0.673±0.008	**0.687**±0.014	0.670±0.027	0.666±0.010
Pneumonia	0.704±0.006	**0.725**±0.004	*0.723*±0.007	0.714±0.005	0.715±0.008	0.707±0.004	0.719±0.008
Respiratory failure	0.800±0.005	**0.816**±0.007	*0.815*±0.003	0.809±0.005	0.810±0.004	0.805±0.004	0.811±0.008
Septicemia (except in labor)	*0.764*±0.005	0.757±0.005	0.760±0.007	0.741±0.005	0.749±0.003	0.763±0.004	**0.774**±0.005
Shock	0.803±0.004	0.803±0.005	0.805±0.006	**0.807**±0.003	0.800±0.008	0.801±0.007	*0.806*±0.007
Average Rank	4.56	4.76	4.32	4.64	3.88	3.84	2

18

B.3 AUPRC of Phenotype Prediction by Disease Label

We summarize the AUPRC score for the phenotype prediction task by disease label in Table 12.

Table 12: The AUPRC score by disease labels with detailed standard deviation.

	Uni-EHR [23]	MMTM [52]	DAFT [9]	MedFuse [10]	DrFuse [13]	GAN-based [53]	DDL-CXR (ours)
Acute renal failure	*0.573*±0.004	0.568±0.007	0.572±0.006	0.565±0.005	0.564±0.005	0.563±0.005	**0.588**±0.008
Acute cerebrovascular disease	0.425±0.010	0.418±0.010	0.419±0.017	*0.434*±0.017	0.399±0.019	**0.446**±0.010	0.416±0.018
Acute myocardial infarction	0.185±0.008	0.192±0.004	0.187±0.009	**0.219**±0.015	*0.209*±0.013	0.171±0.008	0.206±0.020
Cardiac dysrhythmias	0.579±0.006	0.532±0.015	0.548±0.005	0.560±0.012	*0.584*±0.009	0.561±0.008	**0.605**±0.009
Chronic kidney disease	0.515±0.016	0.505±0.011	*0.515*±0.009	0.497±0.009	0.477±0.016	0.501±0.006	**0.538**±0.010
COPD and bronchiectasis	0.319±0.012	0.327±0.021	0.342±0.011	0.344±0.004	**0.405**±0.010	0.372±0.012	*0.382*±0.004
Surgical complications	0.370±0.009	0.379±0.006	*0.385*±0.007	0.381±0.006	0.377±0.003	0.344±0.006	**0.388**±0.004
Conduction disorders	0.276±0.006	0.287±0.007	0.298±0.013	0.286±0.016	*0.632*±0.010	0.609±0.004	**0.633**±0.003
CHF; nonhypertensive	0.593±0.007	0.619±0.019	0.647±0.017	0.631±0.011	*0.661*±0.015	0.652±0.012	**0.682**±0.009
CAD	0.560±0.007	0.540±0.006	0.556±0.008	0.544±0.009	0.581±0.012	*0.590*±0.014	**0.611**±0.010
DM with complications	*0.562*±0.013	**0.569**±0.016	0.552±0.006	0.561±0.012	0.550±0.015	0.552±0.019	0.524±0.007
DM without complication	**0.370**±0.005	0.367±0.013	0.343±0.014	0.356±0.006	*0.369*±0.010	0.352±0.012	0.368±0.008
Disorders of lipid metabolism	*0.594*±0.007	0.576±0.002	0.570±0.009	0.566±0.003	0.584±0.010	0.587±0.008	**0.601**±0.010
Essential hypertension	0.551±0.006	0.519±0.003	0.525±0.005	0.518±0.009	0.502±0.005	*0.554*±0.011	**0.561**±0.011
Fluid and electrolyte disorders	0.655±0.003	*0.664*±0.005	0.662±0.004	0.656±0.008	0.658±0.006	0.662±0.012	**0.672**±0.008
Gastrointestinal hemorrhage	0.180±0.013	0.142±0.014	0.162±0.019	**0.192**±0.014	*0.191*±0.014	0.151±0.016	0.180±0.009
Secondary hypertension	*0.463*±0.012	0.455±0.007	0.452±0.007	0.453±0.013	0.437±0.012	0.451±0.011	**0.484**±0.009
Other liver diseases	0.316±0.013	0.316±0.010	0.341±0.007	0.344±0.008	*0.372*±0.014	0.362±0.018	**0.378**±0.009
Other lower respiratory disease	0.219±0.003	0.209±0.012	0.206±0.008	0.223±0.005	**0.255**±0.011	0.236±0.008	*0.242*±0.007
Other upper respiratory disease	0.166±0.015	0.137±0.009	0.166±0.019	0.202±0.014	**0.274**±0.018	0.196±0.019	*0.234*±0.059
Pleurisy; pneumothorax	0.143±0.005	0.145±0.008	0.159±0.009	0.159±0.013	**0.172**±0.007	*0.171*±0.022	0.166±0.009
Pneumonia	0.412±0.012	**0.437**±0.005	*0.429*±0.008	0.419±0.008	0.406±0.016	0.415±0.010	0.428±0.013
Respiratory failure	0.655±0.009	*0.686*±0.011	0.674±0.004	0.671±0.008	**0.692**±0.005	0.663±0.007	0.669±0.012
Septicemia (except in labor)	*0.585*±0.008	0.573±0.012	0.580±0.014	0.565±0.009	0.562±0.013	0.573±0.009	**0.603**±0.010
Shock	*0.590*±0.002	0.584±0.005	0.582±0.009	**0.592**±0.008	0.572±0.021	0.587±0.011	0.586±0.013
Average Rank	4.4	4.64	4.4	4.24	3.88	4.16	2.28

B.4 Ablation Study for a Reduced Percentage of Data

To investigate the sensitivity to the amount of training data of the proposed model and the main baselines, we conduct experiments with varying training sizes. The results are summarized in Table 13 and Table 14. Results show that DDL-CXR consistently outperforms baseline methods by a large margin, demonstrating its robustness against training data size.

Table 13: Performance of the phenotype classification task with different training sizes. Results are reported in mean±std.

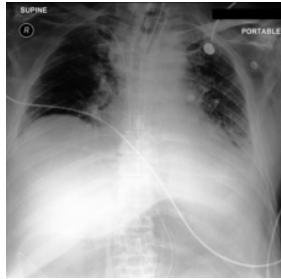
	AUPRC			AUROC		
	100%	75%	50%	100%	75%	50%
Uni-EHR [23]	0.434 ±0.009	0.428 ±0.011	0.419 ±0.010	0.720 ±0.006	0.711 ±0.008	0.706 ±0.006
MMTM [52]	0.430 ±0.005	0.421 ±0.004	0.406 ±0.003	0.715 ±0.003	0.707 ±0.002	0.694 ±0.002
DAFT [9]	0.435 ±0.002	0.422 ±0.004	0.407 ±0.003	0.720 ±0.003	0.709 ±0.002	0.699 ±0.003
MedFuse [10]	0.437 ±0.001	0.420 ±0.004	0.412 ±0.002	0.718 ±0.002	0.707 ±0.003	0.700 ±0.001
DrFuse [13]	0.459 ±0.003	0.442 ±0.005	0.431 ±0.004	0.729 ±0.004	0.717 ±0.005	0.709 ±0.004
DDL-CXR	**0.470** ±0.003	**0.457** ±0.003	**0.433** ±0.011	**0.740** ±0.002	**0.734** ±0.002	**0.715** ±0.005

Table 14: Performance of the mortality prediction task with different training sizes. Results are reported in mean±std.

	AUPRC			AUROC		
	100%	75%	50%	100%	75%	50%
Uni-EHR [23]	0.498 ±0.007	0.437 ±0.011	0.429 ±0.012	0.815 ±0.007	0.791 ±0.005	0.782 ±0.013
MMTM [52]	0.422 ±0.014	0.405 ±0.010	0.399 ±0.012	0.785 ±0.004	0.782 ±0.002	0.775 ±0.005
DAFT [9]	0.448 ±0.004	0.428 ±0.006	0.413 ±0.005	0.800 ±0.003	0.790 ±0.004	0.778 ±0.007
MedFuse [10]	0.443 ±0.009	0.420 ±0.015	0.411 ±0.009	0.793 ±0.003	0.784 ±0.004	0.775 ±0.005
DrFuse [13]	0.460 ±0.004	0.430 ±0.013	0.415 ±0.030	0.773 ±0.008	0.755 ±0.008	0.766 ±0.030
DDL-CXR	**0.523** ±0.011	**0.474** ±0.009	**0.466** ±0.012	**0.822** ±0.009	**0.801** ±0.008	**0.790** ±0.008

19

B.5 Additional Case Studies

We demonstrate additional case studies in Fig. 4 – Fig. 9. We retrieve findings from the radiology reports. The case studies show that DDL-CXR could generate CXR images that align with the disease progression of the individual patient.

(a)*Initial radiology findings*: the chest demonstrates low lung volumes, which results in bronchovascular crowding. Bibasilar opacities may reflect atelectasis There is a probable small left pleural effusion.

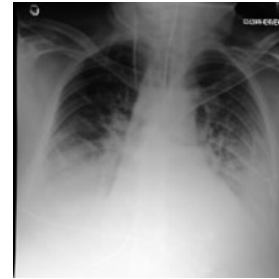
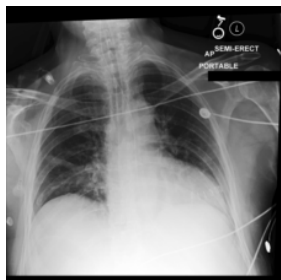
(b) *Radiology findings after 8 hours*: There is interval progression of interstitial pulmonary edema and potential interval increase in bibasal consolidations.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 8 hours.

Figure 4: Case Study of Sample #1

(a) Heart remains enlarged with left ventricular prominence. Interval appearance of linear opacity in the right mid lung which may reflect fluid loculated within the minor fissure or possibly subsegmental atelectasis.

(b) *Radiology findings after 18 hours*: There are fluctuating patchy opacities at the right lung base suggestive of atelectasis. Low volumes with crowding of the pulmonary vasculature.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 18 hours.

Figure 5: Case Study of Sample #2

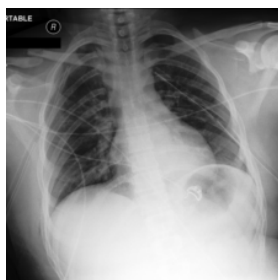
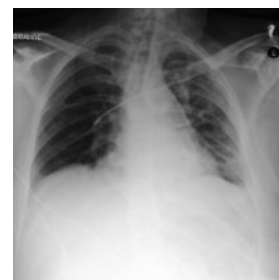
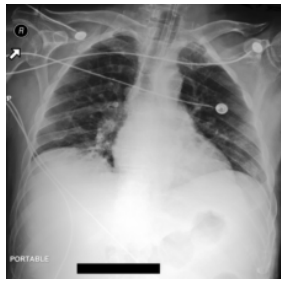
(a) The dense atelectatic streaks in the left mid zone has decreased. The bilateral chest tubes remain in place and there is no evidence of pneumothorax.

(b) *Radiology findings after 13 hours*: No left pneumothorax is appreciated. The hemidiaphragms are now sharp be seen with only mild atelectatic changes at the bases.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 13 hours.

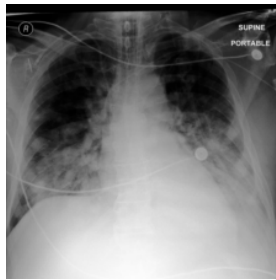
Figure 6: Case Study of Sample #3

20

(a) *Initial radiology findings*:The lung volumes are low. Mild cardiomegaly without pulmonary edema. No pleural effusions.

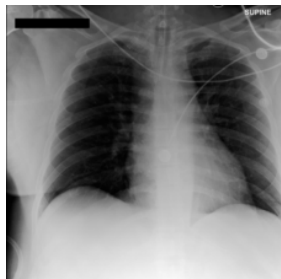
(b) *Radiology findings after 13 hours*: There is mild bibasilar atelectasis. There is no pneumothorax or large pleural effusion.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 13 hours.

Figure 7: Case Study of Sample #5

(a) *Initial radiology findings*: The lungs bilaterally demonstrate severe, extensive rounded nodular densities.

(b) *Radiology findings after 31 hours*: Bilateral nodular parenchymal opacities are unchanged in this patient with known lymphoma. There are likely layering effusions, left greater than right.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 31 hours.

Figure 8: Case Study of Sample #6

(a)*Initial radiology findings*: Heart size is normal. Mediastinal and hilar contours are within normal limits. Pulmonary vasculature is normal.

(b) *Radiology findings after 29 hours*: Small amount of right pleural effusion is present, more conspicuous than on the prior study.

(c) CXR image generated by DDL-CXR given the initial CXR image shown in (a) and the EHR data within the 29 hours.

Figure 9: Case Study of Sample #7

21

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflected the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations in Section 5. We will address these limitations in future work.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

22

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: The related codes for reproducing the main experimental results are submitted in supplementary materials and the results are discussed in Section 4.

Guidelines:

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

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

23

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

Answer: [Yes]

Justification: We submit the codes as supplemental material to the paper submission. The codes will be made publicly available upon paper acceptance. The datasets used in the paper are open-source and publicly accessible.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: The details of data splits and the hyperparameter searching space are illustrated in Section 4.1 and Appendix A.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: The results are reported by taking the average of five runs of model training along with the standard deviations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

24

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

8. **Experiments Compute Resources**

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

Answer: [Yes]

Justification: The details of computer resources for the experiments are given in Appendix A.

Guidelines:

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: We conform in every respect with the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader Impacts**

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

Answer: [Yes]

Justification: The paper discusses both potential societal impacts and negative societal impacts of the work performed in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

25

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

11. **Safeguards**

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

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We properly credited all code, data, and models used in the paper. The licenses are explicitly mentioned.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

26

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects. The paper utilizes de-identified public datasets.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: The study involves only public de-identified datasets and no IRB approvals needed.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

27

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

28