MirrorTD: Constraint-Aware Diffusion Models for Mixed-Type EHR Time Series Generation

Anonymous authors
Paper under double-blind review

Abstract

The generation of synthetic electronic health records (EHRs) data is a critical enabler for ML in healthcare. However, it remains challenging because clinical time series are mixed-type (numerical and categorical), high-dimensional, temporally structured, and subject to constraints such as data validity and patient survival status. In response to these challenges, we propose MirrorTD, a multi-stage score-based diffusion framework that integrates mixed-type Gaussian and discrete diffusion processes with a mirror-mapping variational autoencoder to embed constraints. Specifically, we embed constrained indicators into a continuous latent space via the mirror mapping and utilize an efficient spatio-temporal attention mechanism to capture temporal dynamics and cross-feature dependencies. Experiments on three real-world ICU datasets show that our method produces realistic, diverse, and constraint-compliant synthetic EHRs, advancing synthetic time-series generation for critical-care cohorts.

1 Introduction

The increasing adoption of electronic health records (EHRs) now enables unprecedented opportunities for data-driven healthcare research and the development of machine learning (ML) models for clinical decision support (Goldberger et al., 2000; Miotto et al., 2017; Shickel et al., 2017; Rajkomar et al., 2018). Yet the sensitive nature of patient data and strict privacy regulations often limit data sharing, thereby hindering large-scale ML model training. Consequently, synthetic EHR data generation has emerged as a promising privacy-preserving mechanism that still permits broad application of ML models in healthcare (Nikolentzos et al., 2023). To meet this need, recent literature has explored generative adversarial networks (GANs) (Choi et al., 2017; Baowaly et al., 2019), variational autoencoders (VAEs) (Esteban et al., 2017; Chen et al., 2021), and diffusion-based methods (Li et al., 2023; Ho et al., 2020) for synthesizing realistic EHR time-series data.

Yet several obstacles still limit the realism and downstream utility of synthetic EHR datasets. Foremost is the mixed-type nature of EHR data: numerical features (e.g., lab measurements such as blood pressure) coexist with categorical events (e.g., diagnoses, medications) and require distinct consideration for generative and probabilistic models (Austin et al., 2021; Hoogeboom et al., 2021). Treating mixed-type variables together tangles their continuous and discrete likelihood or loss terms, so training becomes imbalanced and reliable correlations becomes hard to learn (Choi et al., 2017; Li et al., 2023), among several others.

Beyond mixed-type interactions, the second challenge in generating synthetic EHR time series datasets is capturing spatio-temporal dynamics. Clinical trajectories are heterogeneous, vary in length, and are irregularly sampled, producing complex, time-dependent missingness and observation patterns (Rajkomar et al., 2018; Peebles & Xie, 2023). Capturing cross-feature temporal correlations, such as evolving links between continuous laboratory measurements and concurrent categorical events like medications or diagnoses, remains essential for generating plausible patient trajectories and downstream utility (Choi et al., 2017; Esteban et al., 2017). Thus, models that miss local or long-range dependencies inevitably produce spurious co-occurrences and fail to yield clinically plausible synthetic data (Peebles & Xie, 2023; Liu et al., 2023).

Even with accurate temporal dynamics, the yet another challenge is enforcing clinical validity constraints, which remains essential yet difficult for unconstrained generators. These constraints encompass mortal-

ity monotonicity, logical consistency across events, and validity flags observed in ICU records from public datasets such as MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), and eICU (Pollard et al., 2018). Such constraints are diverse and often combinatorial, spanning features and time, and they safeguard clinical plausibility and safety (Liu et al., 2023; Feng et al., 2024). Absent these safeguards, cross-feature correlations may be distorted, causing models trained on synthetic EHRs to diverge from real-world scenarios (Choi et al., 2017; Li et al., 2023).

In this work, we introduce Mirror Time-series Diffusion (MirrorTD), a score-based generative framework that jointly addresses mixed-type feature modeling, spatio-temporal dynamics, and constraint enforcement for EHR time-series data with binary clinical 'flags' (e.g., missingness, mortality). As illustrated in Figure 1, the proposed MirrorTD framework unfolds in the following three stages.

Stage 1 trains a mirror-mapping VAE solely on binary constraint 'flags', yielding compact embeddings that internalize clinical validity requirements. Stage 2 freezes the VAE encoder, concatenates its embeddings with numerical features, and trains a joint diffusion model inspired by Gaussian and discrete formulations (Ho et al., 2020; Lou et al., 2023); the score network employs spatio-temporal attention to capture local transitions and long-range dependencies across categorical and continuous channels. Lastly, the Stage 3 generates categorical trajectories and continuous signals with the diffusion model, after which the VAE decoder deterministically maps synthesized embeddings back to discrete binary 'flags' that satisfy the clinical validity constraints for missingness and mortality.

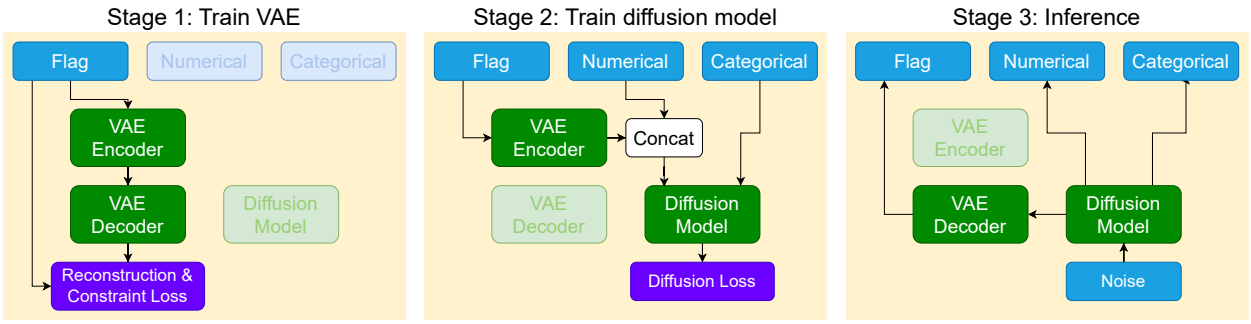

Figure 1: Overview of the MirrorTD framework. *Stage 1:* Train a VAE solely on discrete 'flag' (e.g., missingness, mortality) so its embeddings capture validity constraints. *Stage 2:* Freeze the VAE encoder, append its embeddings to numerical inputs, and train a mixed-type diffusion model whose score network uses spatio-temporal attention across categorical and continuous channels. *Stage 3:* During generation, synthesize categorical data and the continuous vector (numerics + 'flag' embeddings), then decode the embeddings back to discrete binary 'flags' that satisfy the constraints.

Our work fundamentally differs from prior works on diffusion models for time-series health domains, e.g., TimeDiff (Tian et al., 2023) in two key aspects. Firstly, we exploit a novel mirror-mapping VAE to embeds constrained indicators into a continuous latent space and using such an embedding to adhere to clinical constraints for generating novel samples. Secondly, prior works such as TimeDiff relies on a Recurrent Neural Network (RNN) backbone for learning spatio-temporal correlation while, we employ a factorized spatio-temporal transformer architecture which is more expressive in capturing both long-range temporal dynamics and complex cross-feature dependencies.

Effective realization of MirrorTD hinges on satisfying three desiderata formalized in Section 2: preserving constraint 'flags', modeling spatio-temporal structure efficiently, and enabling conditional generation (e.g., for a target disease). Section 3 details how specialized loss functions for the constraint 'flags', factorized transformer attention, and classifier-guided sampling satisfy these desiderata. Section 4 shows that MirrorTD produces realistic, diverse, and constraint-compliant synthetic EHR cohorts, highlighting both clinical plausibility metrics and downstream mortality prediction performance relative to diffusion and VAE

baselines. Section 5 discusses details of prior methods and the key innovations of MirrorTD compared with these methods, and Section 6 summarizes key takeaways.

2 Problem Description

This section formalizes description of the problem. Below, we first define notations followed by the desiderata that any generative mechanism for the EHR time-series dataset must satisfy.

Notation. Scalars, vectors, and matrices use plain letters, lowercase bold, and uppercase bold, respectively. For $N \in \mathbb{N}$, let $[N] = \{1, 2, \ldots, N\}$. The mixed-type dataset is denoted by $\mathcal{D}$, and each record $\mathbf{x} \in \mathcal{D}$ comprises four L-length time-series components:

1. $\mathbf{x}_{\text{num}} \in \mathbb{R}^{d_{\text{num}} \times L}$: Numerical features often arising from ICU measurements, e.g., blood pressure and heart rate.
2. $\mathbf{x}_{\text{cat}} \in \mathbb{R}^{d_{\text{cat}} \times L}$: Categorical features covering patient attributes such as disease indicators and multi-class demographics.
3. $\mathbf{x}_{\text{nan}} \in \mathbb{R}^{d_{\text{nan}} \times L}$: Validity flags, where $\mathbf{x}^{i,l}_{\text{nan}} = 1$ iff the associated numerical feature $\mathbf{x}^{i,l}_{\text{num}}$ is missing; in MIMIC-III, $d_{\text{nan}} = d_{\text{num}}$.
4. $\mathbf{x}_{\text{mort}} \in \mathbb{R}^{1 \times L}$: Mortality flags, with $\mathbf{x}^{l}_{\text{mort}} = 1$ denoting death by time index l.

2.1 Desiderata.

The goal of generative models is to learn a framework that produces synthetic record $\mathbf{x}$ indistinguishable from those in the EHR dataset $\mathcal{D}$. Concretely, we seek to train a neural model that allows to sample from $p(\mathbf{x})$ while meeting the clinical desiderata mentioned below.

Desideratum-1: *Preserving validity and mortality constraints.* EHR time-series records have validity ($\mathbf{x}_{\text{nan}}$) and mortality ($\mathbf{x}_{\text{mort}}$) flags which must satisfy the following constraints for any synthesized record $\mathbf{x}$:

- Monotone mortality flags, i.e., for any $l \in [L-1]$, $\mathbf{x}^{l}_{\text{mort}} = 1$ implies $\mathbf{x}^{l+1}_{\text{mort}} = 1$;
- validity flags marks all continuous measurements invalid once the patient deceases, i.e., for any $l \in [L]$ and $i \in [d_{\text{nan}}]$, $\mathbf{x}^{l}_{\text{mort}} = 1$ implies $\mathbf{x}^{i,l}_{\text{nan}} = 1$.

Desideratum-2: *Efficient spatio-temporal modeling.* Naively modeling spatio-temporal correlations with deep networks incurs quadratic complexity in both the number of features (d) and the sequence length (L), i.e., $\mathcal{O}(d^2L^2)$. This effective d can exceed the count of raw features because categorical fields are embedded into higher-dimensional representations. Thus, it is desirable to reduce the computational burden without compromising the fidelity of the EHR generative model.

Desideratum-3: *Conditional generation for EHR records.* Rare cohorts, defined by specific comorbidities, demographics, or interventions, are sparse in the dataset, yet applying ML models to these populations remains highly valuable. Consequently, the ability to conditionally generate EHR time series, i.e., to sample $p(\mathbf{x} \mid c)$ for a cohort condition c, is essential for synthetic data generation.

3 Method

This section introduces MirrorTD, our diffusion-based framework for generating EHR time-series datasets. We begin in Section 3.1 by presenting the diffusion model for mixed-type EHR time series (as described in Section 2) along with its training objective. The remaining subsections elucidate how the proposed MirrorTD fulfills the desiderata given in Section 2. Concretely, Section 3.2 enforces validity/mortality constraints via mirror diffusion, Section 3.3 develops an efficient spatio-temporal transformer to capture

long-range dependencies, and Section 3.4 details the conditional sampling mechanism. The overall schematic of the proposed MirrorTD framework is illustrated in Figure 1.

Let (E, D) denote the VAE encoder-decoder pair in the mirror diffusion, then our framework can be partitioned into the following three stages (as also illustrated in Figure 1):

- **Stage 1: Training the latent encoder–decoder module.** Using **only the flag features** $\mathbf{x}_{\text{mort}}$ and $\mathbf{x}_{\text{nan}}$, we train the encoder and decoder by minimize the loss $\mathcal{L}_{\text{mirror}}$.
- **Stage 2: Diffusion model training.** We freeze the weights of the autoencoder and obtain the flag embedding $\mathbf{z}_{\text{flag}} = E(\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}})$, and concatenate $\mathbf{z}_{\text{flag}}$ with $\mathbf{x}_{\text{num}}$ to get the features for Gaussian diffusion $\mathbf{x}_{\text{g}}$. We then train the mixed diffusion model with $\mathbf{x}_{\text{g}}$ and $\mathbf{x}_{\text{cat}}$.
- **Stage 3: Inference.** We use the diffusion model to get $\mathbf{x}_{\text{g}}$ and $\mathbf{x}_{\text{cat}}$. $\mathbf{x}_{\text{g}}$ is further partitioned into the flag embedding $\mathbf{z}_{\text{flag}}$ and the numerical features $\mathbf{x}_{\text{num}}$. We finally get the flag features $(\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}}) = D(\mathbf{z}_{\text{flag}})$.

3.1 Diffusion model for mixed-type EHR

We utilize coupled Gaussian and discrete diffusion processes to build the generative model, adding noise during the forward diffusion step and removing it during the backward pass. Let $\mathbf{x}^{(0)}$ be the clean data. In the forward process, the distribution of each feature with added noise is:

$$q(\mathbf{x}^{\cdot,l,(1:T)}|\mathbf{x}^{\cdot,l,(0)}) = \prod_{t=1}^{T} q(\mathbf{x}^{\cdot,l,(t)}|\mathbf{x}^{\cdot,l,(t-1)}).$$

To incorporate both continuous- and discrete-valued time series, we employ a "mixed sequence diffusion" scheme that injects Gaussian noise inspired by denoising diffusion probabilistic models (Ho et al., 2020) and categorical noise modeled via multinomial diffusion (Hoogeboom et al., 2021) during the forward pass. Concretely, the overall score function $\mathbf{s}^{t,\boldsymbol{\theta}}(\mathbf{x})$ comprises two components: the Gaussian term $\mathbf{s}_{\text{g}}^{t,\boldsymbol{\theta}}(\mathbf{x})$ and the categorical term $\mathbf{s}_{\text{cat}}^{t,\boldsymbol{\theta}}(\mathbf{x})$. Let $\mathcal{L}^{\text{g}}(\boldsymbol{\theta})$ and $\mathcal{L}^{\text{cat}}(\boldsymbol{\theta})$ be the loss functions for the Gaussian diffusion and multinomial diffusion, respectively, and let $\lambda > 0$ be a weighting parameter. Then the total training loss is

$$\mathcal{L}(\boldsymbol{\theta}) = \lambda \cdot \mathcal{L}^{\text{g}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{cat}}(\boldsymbol{\theta}).$$

The reader can refer to Appendix A.2 for more details of the diffusion model.

3.2 Mirror Diffusion for Constraint Enforcement

To address Desideratum-1 from Section 2, we draw inspiration from mirror diffusion (Liu et al., 2023; Feng et al., 2024) and learn a continuous latent representation for the discrete structured constraints. Both mortality and data-validity constraints can be jointly described by the set C:

$$C = \big\{(\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}}) \in \{0,1\}^{(d_{\text{nan}}+1)\times L} \big| \mathbf{x}_{\text{mort}}^{l+1} \geq \mathbf{x}_{\text{mort}}^{l}, \mathbf{x}_{\text{nan}}^{i,l} \geq \mathbf{x}_{\text{mort}}^{l}, \forall l \in [L], \forall i \in [d_{\text{nan}}]\big\}.$$

To handle the discrete, non-convex nature of C, we relax the constraints to a continuous domain via mirror diffusion. Specifically, we consider the convex hull of C, denoted C^+, and restrict the diffusion model outputs to the convex set C^+. The mirror diffusion module relies on an encoder-decoder pair (E, D) with mappings $E : C^+ \to \mathbb{R}^d$ and $D : \mathbb{R}^d \to [0,1]^{(d_{\text{nan}}+1)\times L}$. The reconstruction loss enforces faithful recovery of the input domain, i.e., both flag features:

$$\begin{aligned} &\mathcal{L}_{\text{recon}} = \mathbb{E}_{(\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}}) \in C^+} \big[\text{BCE}\big((\mathbf{y}_{\text{mort}}, \mathbf{y}_{\text{nan}}), (\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}})\big)\big], \\ &\text{where} \quad (\mathbf{y}_{\text{mort}}, \mathbf{y}_{\text{nan}}) = D\big(E(\mathbf{x}_{\text{mort}}, \mathbf{x}_{\text{nan}})\big). \end{aligned}$$

Here BCE is the Binary Cross Entropy loss:

$$\text{BCE}(y, x) = -\frac{1}{d} \sum_{i=1}^{d} \left[x_i \log y_i + (1 - x_i) \log(1 - y_i)\right].$$

A complementary constraint loss is also applied to penalize the violations of C^+:

$$\mathcal{L}_{\text{const}} = \sum_{l=1}^{L-1} (\mathbf{y}^l_{\text{mort}} - \mathbf{y}^{l+1}_{\text{mort}})_+ + \sum_{i=1}^{d_{\text{nan}}} \sum_{l=1}^{L} (\mathbf{y}^l_{\text{mort}} - \mathbf{y}^{i,l}_{\text{nan}})_+,$$

where $(z)_+ = \max\{z, 0\}$. Minimizing $\mathcal{L}_{\text{const}}$ keeps decoder outputs within C^+, and coupled with the loss function $\mathcal{L}_{\text{recon}}$ keeps them close to the original constraint set C because the inputs lie in the set C. With weighting parameter $\lambda_{\text{m}} > 0$, the total loss is $\mathcal{L}_{\text{mirror}} = \mathcal{L}_{recon} + \lambda_{\text{m}} \cdot \mathcal{L}_{\text{const}}$.

3.3 Spatio-Temporal Transformer

This section addresses Desideratum 2 from Section 2. For the diffusion model over mixed-type EHR time series, the score function $\mathbf{s}_{t,\boldsymbol{\theta}}(\mathbf{x}_t, t)$ must capture two distinct dependencies: (1) spatial correlations that reflect synchronous relationships between heterogeneous clinical features (e.g., blood pressure and heart rate) and (2) temporal correlations that track the longitudinal evolution of a patient's state across irregular intervals. To model these dependencies efficiently, we use the factorized spatio-temporal attention mechanism introduced by Tashiro et al. (2021) Unlike traditional 1D convolutions or flat self-attention, the factorized approach splits the attention operation to avoid the quadratic complexity in the joint product of sequence length L and feature count d.

Factorized Spatio-Temporal Attention (FSTA). Let $\mathbf{Z} \in \mathbb{R}^{d \times L \times h}$ be the embedded features, where h is the dimension of the embedding. The FSTA module operates as follows:

1. *Spatial attention (i.e., intra-timestep):* We first apply the spatial attention $\text{Attn}_{\text{s},\cdot}$ across the feature dimension for each time step $l \in [L]$. This step models the conditional dependence between disparate clinical markers (e.g., lab results and medication flags) at a specific point in time:

$$\text{Attn}_{\text{s},l} = \text{Softmax}\Big(\frac{\mathbf{Q}_l \mathbf{K}_l^\top}{\sqrt{d_k}}\Big)\mathbf{V}_l, \text{ where } (\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l) = \mathbf{Z}_{:,l,:} \cdot (\mathbf{W}_{\text{s},Q}, \mathbf{W}_{\text{s},K}, \mathbf{W}_{\text{s},V}).$$

   $\mathbf{W}_{\text{s},Q}, \mathbf{W}_{\text{s},K}, \mathbf{W}_{\text{s},V} \in \mathbb{R}^{h \times d_k}$ are trainable parameters of the spatial attention.

2. *Temporal attention (i.e., inter-timestep):* Following the spatial update, we apply the temporal attention $\text{Attn}_{\text{t},\cdot}$ across the temporal axis. For each feature index $i \in [d]$, the model computes attention weights across the L time steps. This allows the score estimator to identify long-range dependencies and "smooth" noise estimations based on historical trajectories:

$$\text{Attn}_{\text{t},i} = \text{Softmax}\Big(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_k}}\Big)\mathbf{V}_i, \text{ where } (\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \mathbf{Z}_{i,:,:} \cdot (\mathbf{W}_{\text{t},Q}, \mathbf{W}_{\text{t},K}, \mathbf{W}_{\text{t},V}).$$

   $\mathbf{W}_{\text{t},Q}, \mathbf{W}_{\text{t},K}, \mathbf{W}_{\text{t},V} \in \mathbb{R}^{h \times d_k}$ are trainable parameters of the temporal attention.

Following Peebles & Xie (2023), we use the adaptive layer normalization (AdaLN) instead of the standard layer normalization in the spatio-temporal transformer to process the diffusion step t. See Appendix A.2 for details. The integration of FSTA into the diffusion backbone serves as a structural prior for the score function. By alternating between spatial and temporal pathways, the model effectively learns the joint distribution $p(\mathbf{x}_0)$.

Handling mixed-type data. This architecture allows the model to handle mixed-type data, where categorical features are embedded via `nn.Embedding` and continuous features via scalar projection. By allowing the attention mechanism to learn a shared geometric space where cross-type interactions are quantified through attention scores. This enables the generative process to maintain logical consistency (e.g., ensuring that a "hypoglycemia" categorical flag is temporally aligned with a decrease in "blood glucose" continuous values).

Complexity and Scalability. By factorizing the attention into separate spatial and temporal attention blocks, the computational complexity of STA is $O(dL^2 + Ld^2)$ rather than a joint $O(d^2L^2)$. In experiments with MIMIC-III/IV and eICU, we have $L \geq d \geq 9$, so the factorized attention mechanism accelerates the spatio-temporal attention by at least 4.5 times. Therefore, the STA module remains computationally tractable for extended EHR records while providing the high-capacity modeling.

3.4 Classifier-Free Guidance Enables Conditional Generation

This section shows how Desideratum-3 from Section 2 can be addressed using classifier-free guidance (CFG, Ho & Salimans 2022) for conditional generation. Recall that c denotes the desired cohort condition, e.g., indicators of mortality, age group, or disease name.

Compared with the unconditional generation in Section 3.1, the modified score network $\mathbf{s}^{\boldsymbol{\theta},t}(\mathbf{x}, c)$ can be trained with the condition c as an additional argument. As long as the conditional information can be ingested by the score network, the other components of the proposed MirrorTD remain compatible, thus enabling conditional generation for EHR time-series datasets. Specifically, the inference process is guided to sample data with high $p(\mathbf{x} \mid c)$ by replacing the unconditional score function $\mathbf{s}^{\boldsymbol{\theta},t}(\mathbf{x})$ with the guided score function $\hat{\mathbf{s}}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}, c)$ defined as

$$\hat{\mathbf{s}}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}, c) = \beta \cdot \mathbf{s}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}, c) + (1 - \beta) \cdot \mathbf{s}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}, \varnothing),$$

where $\beta > 0$ is the hyperparameter controlling the strength of classifier-free guidance, and $\varnothing$ represents a "null" token corresponding to unconditional sampling. It is worth noting that the same score network can be trained for both scenarios, i.e., with and without the conditional information.

4 Experiments

This section empirically evaluates MirrorTD on three publicly available ICU datasets (MIMIC-III, MIMIC-IV, and eICU) against five generative baselines. Section 4.1 describes the datasets and pre-processing pipeline, and Section 4.2 outlines the baselines. Sections 4.3–4.5 assess distributional realism: Section 4.3 compares per-feature prevalence to verify that synthetic samples match the marginal statistics of real EHRs; Section 4.4 trains a GRU classifier to test whether real and synthetic samples are jointly indistinguishable; and Section 4.5 trains a GRU regressor on synthetic data to test whether the learned conditional dynamics transfer to real test sequences. Section 4.6 then compares the temporal patterns of synthetic and real trajectories via Markov state transition matrices. Sections 4.7–4.9 evaluate downstream and structural properties: Section 4.7 reports in-hospital mortality prediction accuracy under the TSTR/TRTR protocol; Section 4.8 evaluates privacy through four complementary attacks (NNAA, MIA AUC, MIR, AIR); and Section 4.9 reports clinical-constraint violation rates for the two hard constraints introduced in Section 2. Section 4.10 evaluates conditional generation via classifier-free guidance under two conditioning schemes.

Across all axes, MirrorTD attains the lowest mean discriminative score on every dataset, with a large margin on eICU (0.032 vs. 0.263 for TimeDiff) and MIMIC-IV (0.161 vs. 0.245), and comparable performance on MIMIC-III (0.189 vs. 0.192, within one standard deviation). It also achieves the lowest Frobenius distance to real Markov transitions (0.227 vs. 0.430 on MIMIC-III) and reduces hard-constraint violations by over an order of magnitude (e.g., 4.1% vs. 77.9% on MIMIC-III for mortality monotonicity), while remaining competitive on downstream mortality prediction and matching TimeDiff on privacy.

4.1 Dataset and Pre-processing Details

We evaluate MirrorTD on three public datasets—MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), and eICU (Pollard et al., 2018)—restricting attention to their ICU cohorts. Following Tian et al. (2023), we preprocess each dataset to obtain training tensors containing $\mathbf{x}_{\text{num}}$, $\mathbf{x}_{\text{cat}}$, $\mathbf{x}_{\text{nan}}$, and $\mathbf{x}_{\text{mort}}$. Every dataset provides numerical measurements and their associated flags: $\mathbf{x}_{\text{num}}$ captures ICU vitals (e.g., blood pressure, heart rate, oxygen saturation) and $\mathbf{x}_{\text{nan}}$ denotes the missing-value indicator for each numerical channel. MIMIC-IV is the only benchmark with non-empty categorical features $\mathbf{x}_{\text{cat}}$, which encode diagnoses,

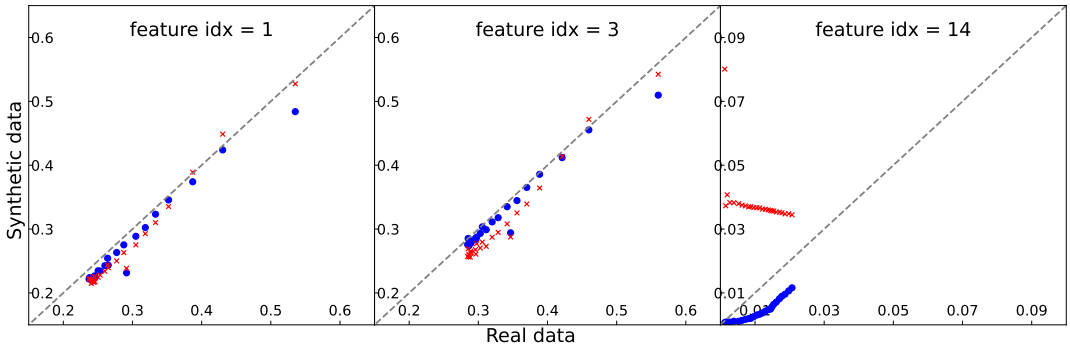

Figure 2: Prevalence comparison of MirrorTD (blue) and TimeDiff (red) on MIMIC-III. Each point corresponds to a feature's empirical mean (NaN flags `idx` $= 1, 3$ and mortality flag `idx` $= 14$); alignment with the diagonal indicates perfect agreement with real data. MirrorTD remains close to the diagonal, highlighting superior prevalence fidelity relative to the TimeDiff baseline.

medications, and patient age groups. Dataset-level statistics appear in Table 5, with further preprocessing details in Appendix A.1.

4.2 Baseline Approaches

We benchmark the MirrorTD framework against three representative generative models: GT-GAN (Jeon et al., 2022), DSPD/CSPD (Biloš et al., 2023), and TimeDiff (Tian et al., 2023).

- GT-GAN adopts a GAN-style architecture in which a neural ODE evolves the latent state.
- DSPD (discrete-time) and CSPD (continuous-time) treat the entire sequence as a function and train diffusion models directly in that functional space.
- TimeDiff, the strongest prior baseline, also employs mixed diffusion but handles mortality/validity flags as ordinary categorical variables.

For each dataset and each baseline approach, we generate 20000 samples for evaluation.

4.3 Prevalence Fidelity Analysis

Prevalence evaluates marginal fidelity by measuring the empirical mean of each feature across the cohort at every time step. For $l \in [L]$ and feature i, we compute $\text{Prev}_{i,l} = \frac{1}{N}\sum_{n=1}^{N} x_{i,l}^{(n)}$ on both the real and synthetic datasets. When plotted against one another (synthetic on the y-axis, real on the x-axis), perfectly matched prevalence lies along the $y = x$ diagonal.

Figure 2 provides prevalence results for both MirrorTD and TimeDiff approaches on MIMIC-III (additional datasets appear in Appendix B). MirrorTD tracks the NaN-flag prevalence (`idx` $= 1, 3$) much more closely, whereas TimeDiff systematically underestimates missingness, yielding overly complete synthetic records. For the mortality flag (`idx` $= 14$), the rarest feature, TimeDiff overshoots the positive rate, while MirrorTD reproduces the expected monotonic rise seen in real ICU trajectories because it stays much closer to the diagonal. These results suggest that the proposed MirrorTD faithfully preserves the prevalence of critical features in synthetic EHR time series, which is essential for downstream utility and clinical plausibility.

4.4 Discriminative Score Analysis

Following Jeon et al. (2022); Tian et al. (2023), we train a GRU classifier to distinguish real from synthetic records and report discriminative score which is set to $|0.5 - \text{Accuracy}|$. Scores near zero imply that the classifier cannot separate the two distributions, indicating high realism.

Table 1 shows that MirrorTD attains the lowest discriminative scores on eICU, MIMIC-III, and MIMIC-IV, outperforming TimeDiff and other baselines whose scores remain close to 0.5 (i.e., the classifier easily detects them). Across all three datasets, MirrorTD consistently surpasses TimeDiff as well as the remaining generative baselines. For example, on eICU the discriminative score drops from 0.263 for TimeDiff to 0.032 for MirrorTD, an absolute improvement of roughly **23**%, underscoring the improved realism of samples drawn from our framework.

Table 1: Predictive and discriminative scores of MirrorTD and baselines (lower is better). For further details on results and metrics, please see Sections 4.4 and 4.5

Metric	Method	eICU	MIMIC-III	MIMIC-IV
Discriminative Score (↓)	*Real Data*	0.051 ± 0.035	0.012 ± 0.011	0.007 ± 0.006
	TimeDiff	0.263 ± 0.090	0.192 ± 0.029	0.245 ± 0.073
	DSPD-GP	0.442 ± 0.040	0.495 ± 0.007	0.494 ± 0.003
	DSPD-OU	0.471 ± 0.030	0.497 ± 0.001	0.494 ± 0.005
	CSPD-GP	0.427 ± 0.046	0.498 ± 0.001	0.483 ± 0.028
	CSPD-OU	0.439 ± 0.035	0.490 ± 0.004	0.497 ± 0.002
	GT-GAN	0.415 ± 0.059	0.493 ± 0.002	0.447 ± 0.032
	MirrorTD	**0.032 ± 0.035**	**0.189 ± 0.083**	**0.161 ± 0.102**
Predictive Score (↓)	*Real Data*	0.334 ± 0.043	0.441 ± 0.003	0.303 ± 0.002
	TimeDiff	0.339 ± 0.016	**0.448 ± 0.002**	0.328 ± 0.011
	DSPD-GP	0.559 ± 0.036	0.798 ± 0.036	0.532 ± 0.031
	DSPD-OU	0.592 ± 0.039	0.720 ± 0.048	0.736 ± 0.072
	CSPD-GP	0.411 ± 0.056	0.783 ± 0.005	0.534 ± 0.041
	CSPD-OU	0.569 ± 0.048	0.675 ± 0.024	0.610 ± 0.065
	GT-GAN	0.492 ± 0.051	0.633 ± 0.029	0.412 ± 0.017
	MirrorTD	**0.314 ± 0.023**	0.455 ± 0.003	**0.321 ± 0.007**

4.5 Predictive Score Analysis

To assess temporal fidelity we train a GRU regressor exclusively on synthetic data to predict the next time step and evaluate Mean Absolute Error (MAE) on real test sequences. Lower MAE indicates that the synthetic data preserves conditional dynamics needed for forecasting.

The lower block of Table 1 reports predictive scores. MirrorTD attains the lowest MAE on eICU (0.314 vs. 0.339 for TimeDiff, a 7.4% relative improvement) and on MIMIC-IV (0.321 vs. 0.328, 2.1%), and is essentially tied with TimeDiff on MIMIC-III (0.455 vs. 0.448). The remaining baselines lag substantially across all three datasets (e.g., GT-GAN ranges from 0.412 to 0.633). These results suggest that the mirror-diffusion design preserves the conditional dynamics needed for downstream forecasting on real trajectories.

4.6 Probing Temporal Dynamics through Markov State Transitions

The Markov state model (MSM) metric probes dynamical aspects of the generated time series by clustering each time step of every EHR record into 10 states. We use k-means clustering to discover these states and then estimate transition matrices from the resulting state sequences. We visualize the matrices and quantify deviations using the Frobenius norm between synthetic and real transitions.

Figure 3 and the associated error heatmaps show that MirrorTD more faithfully recreates the real transition structure on MIMIC-III than TimeDiff, particularly for less probable transitions concentrated in the second column of the matrix. Appendix B reports tabled Frobenius errors, where MirrorTD consistently yields the lowest discrepancies across datasets, indicating superior modeling of local clinical dynamics.

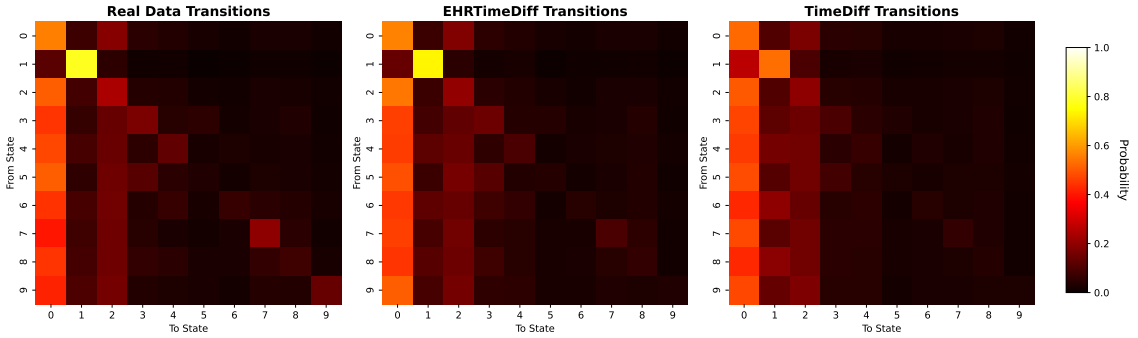

Figure 3: Visualization of MSM transition matrices for the real dataset, MirrorTD, and TimeDiff on MIMIC-III. MirrorTD exhibits substantially higher similarity to the real transitions than TimeDiff.

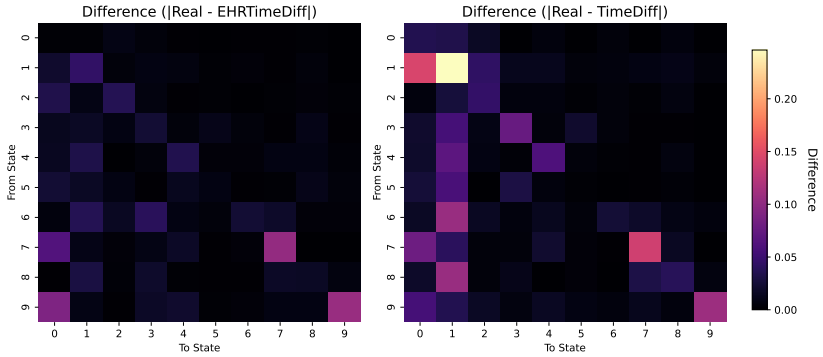

Figure 4: Difference between the synthetic and real MSM transition matrices. Lighter colors for MirrorTD indicate transitions closer to the ground truth than TimeDiff.

4.7 In-Hospital Mortality Prediction

To measure downstream utility we follow the TSTR/TRTR protocol of Tian et al. (2023). Classifiers are trained either on synthetic data (TSTR) or on real data (TRTR) and are always evaluated on held-out real patients for the in-hospital mortality task. We also report a *Real-data* baseline in which both training and evaluation use real records, providing an approximate upper bound for each classifier.

Table 2 summarizes accuracy for five classifiers across these three regimes. In every case, models trained on MirrorTD samples match or exceed the performance of their counterparts trained on TimeDiff data. This indicates that MirrorTD preserves clinically actionable relationships and supports effective synthetic-to-real transfer.

Table 2: Comparison of classifiers trained on synthetic data generated by MirrorTD versus TimeDiff for downstream mortality prediction on MIMIC-III; *Real-data* denotes training on real data.

Predictor	XGBoost	Random Forest	AdaBoost	LR (ℓ_1)	LDA
TimeDiff	0.899 ± 0.027	0.839 ± 0.019	**0.913 ± 0.010**	0.886 ± 0.014	0.916 ± 0.005
MirrorTD	**0.906 ± 0.016**	**0.912 ± 0.014**	0.905 ± 0.007	**0.903 ± 0.018**	**0.917 ± 0.009**
Real-data	0.942 ± 0.006	0.920 ± 0.004	0.939 ± 0.007	0.925 ± 0.007	0.913 ± 0.006

Table 3: Privacy evaluation following the protocol of Tian et al. (2023). **MIA AUC** ≈ 0.5: membership inference attacker performs at chance. **NNAA loss** ≈ 0: model treats train and test symmetrically, indicating no differential memorization. **MIR F1** $\ll 2/3$: upper bound corresponds to a degenerate attacker flagging all samples as members. **AIR leakage** ≤ 0: no additional sensitive attribute information leaked relative to the non-member baseline. MirrorTD and TimeDiff achieve comparable privacy; MIA AUC confirms neither model memorizes training data.

Dataset	Model	NNAA loss ↓ (ideal: 0)	MIA AUC (ideal: 0.5)	MIR F1[‡] ↓ (ref: ≤2/3)	AIR_{Mort} leakage (ref: ≤0)
MIMIC-III	TimeDiff	0.003	0.518	0.173	−0.077
	MirrorTD	0.007	0.519	0.178	−0.033
MIMIC-IV	TimeDiff	0.001	0.494	0.027	N/A[†]
	MirrorTD	0.001	0.492	0.033	N/A[†]
eICU	TimeDiff	0.013	0.438	0.022	−0.052
	MirrorTD	0.016	0.442	0.053	−0.114

[‡] Fixed threshold τ=0.08 (MIMIC-III/IV), τ=0.005 (eICU), following Tian et al. (2023). Upper bound 2/3 corresponds to a no-skill attacker that flags all samples as members.
[†] Synthetic mortality prevalence ≈ 0; single-class collapse.

4.8 Privacy Analysis

We evaluate the privacy preservation of MirrorTD using four complementary metrics: Nearest-Neighbor Adversarial Accuracy (NNAA, Yan et al. 2022), Membership Inference Attack AUC (MIA, Yan et al. 2022), Membership Inference Risk (MIR, Yan et al. 2022), and Attribute Inference Risk (AIR, Yan et al. 2022).

Nearest-Neighbor Adversarial Accuracy (NNAA). NNAA measures whether an adversary can distinguish synthetic from real samples using nearest-neighbor distances. For each split (train or test), we compute the fraction of samples whose closest neighbor from the synthetic set is farther than their closest real neighbor (AA score). The *privacy loss* is defined as $\text{AA}_{\text{test}} - \text{AA}_{\text{train}}$; a value near zero indicates the model does not memorize training data beyond what is captured by the test distribution. We use $n = 5{,}000$ samples per split.

Membership Inference Attack (MIA) AUC. MIA measures whether an attacker can determine, for a given real record, whether it was part of the training set. Following Yan et al. (2022), we use nearest-neighbor distances to the synthetic set as a proxy score and report the AUC of the resulting binary classifier. An AUC of 0.5 indicates that the attacker performs at chance level, i.e., the synthetic data provides no signal about training membership.

Membership Inference Risk (MIR). MIR quantifies how well an attacker can identify training members by finding near-duplicate synthetic records. Following Yan et al. (2022) and the TimeDiff evaluation protocol Tian et al. (2023), we normalize per-split nearest-synthetic distances to $[0, 1]$ and apply a fixed dataset-specific threshold τ: $\tau = 0.08$ for MIMIC-III and MIMIC-IV, and $\tau = 0.005$ for eICU. F1 score of the resulting binary classifier (positive = training member) is reported; lower F1 indicates stronger privacy.

Attribute Inference Risk (AIR). AIR measures whether synthetic data enables an adversary to infer sensitive attributes of real training records. We train an ensemble of XGBoost and logistic regression classifiers on synthetic data and evaluate on real training records ($\text{AIR}_{\text{synth}}$). As a baseline, we repeat with classifiers trained on real held-out test data ($\text{AIR}_{\text{baseline}}$), which represents the expected performance without any additional leakage from synthetic data. The leakage is defined as $\Delta\text{AIR} = \text{AIR}_{\text{synth}} - \text{AIR}_{\text{baseline}}$; negative or near-zero values indicate that synthetic data provides no additional information about training members beyond what is already observable from non-member data. We evaluate Mortality and HR miss-

ingness flag on MIMIC-III, Mortality and age group on MIMIC-IV, and Mortality on eICU. All scores are reported as AUC, averaged over 3 random subsampling repetitions with $n = 3{,}000$ samples per split.

Results. Table 3 summarizes the privacy metrics for MirrorTD and TimeDiff across all three datasets. The MIA AUC values of 0.49–0.52 across all dataset–model combinations confirm that a state-of-the-art membership inference attacker performs at chance level, providing the strongest direct evidence of privacy preservation. MIR F1 scores are well below the no-leakage upper bound of 2/3, NNAA privacy losses are near zero, and all AIR leakage values are negative or negligible. The two models achieve comparable privacy across all metrics, with MirrorTD exhibiting marginally higher distance-based values consistent with its improved generation fidelity—a well-known tradeoff between distributional closeness and privacy risk Yale et al. (2020). Crucially, MIA AUC confirms that this fidelity improvement does not translate to any actual memorization advantage for an attacker. The N/A entries for MIMIC-IV reflect a known generation-quality limitation in both models (near-zero synthetic mortality prevalence), reported separately in the fidelity analysis.

4.9 Clinical Constraint Satisfaction

A key desideratum for synthetic EHR time series is that generated samples respect the temporal logic of clinical events. We evaluate two hard constraints, C1 that stands for monotone mortality falg, and C2 that stands for NaN flags after death, grounded in clinical reality.

Both constraints are automatically satisfied by the real training data (zero C1 violations; the rare C2 violations in real data arise from a small number of recording artifacts). MirrorTD enforces C1 and C2 at the continuous level: the mirror-mapping decoder projects the diffusion output onto the relaxed feasible set C^+ before any sample is returned. Residual violations arise only from the binarisation step that rounds continuous probabilities to binary flags; these can be further reduced by threshold calibration or a single feasibility-projection step. TimeDiff Tian et al. (2023), by contrast, applies no such projection and must rely on the model implicitly learning the constraints from data alone.

Results. Table 4 reports violation rates computed over generated patients who have at least one positive mortality timestep.

On MIMIC-III, TimeDiff violates C1 in 77.9% of deceased-patient sequences and C2 in 27.8%, indicating that the model frequently "resurrects" patients or continues recording measurements after death. MirrorTD reduces these rates to 4.1% and 0.9%, respectively. We note that the residual violations stem from borderline decoder outputs near the binarisation threshold ($\hat{m}_t > 0.5$): because the mirror-mapping decoder produces continuous probabilities that are subsequently rounded to binary flags, outputs that fall marginally on the wrong side of 0.5 can introduce isolated single-timestep reversals. Importantly, these cases do not reflect a structural failure of the constraint mechanism, as the underlying continuous outputs already approximate the feasible set $\mathcal{C}^+$ and could be further reduced by threshold calibration.

On MIMIC-IV, MirrorTD generates no deceased patients ($N_{\text{dead}} = 0$) out of 20,480 synthetic records, and therefore incurs zero violations by default. This outcome reflects a broader difficulty shared by both models in reproducing the rare mortality class in MIMIC-IV: the real dataset has a substantially lower mortality prevalence compared to MIMIC-III, and the diffusion model's tendency to concentrate probability mass on the majority class suppresses the generation of positive mortality flags. We note that this limitation can be directly addressed through conditional generation; indeed, when guided with a mortality-based conditioning scheme, MirrorTD successfully generates decedent cohorts with mortality rates of 81–85%, demonstrating that it can produce constraint-compliant deceased-patient records when explicitly guided to do so. TimeDiff, by contrast, generates 6,169 deceased patients on MIMIC-IV but violates C1 in 55.0% and C2 in 95.3% of these sequences, underscoring that higher generation volume does not translate to constraint compliance.

Taken together, these results highlight that diffusion models cannot reliably learn hard temporal constraints from data alone without an explicit enforcement mechanism. The mirror-mapping decoder in MirrorTD provides a principled, *training-free* solution that ensures clinical validity at generation time, independent of class prevalence or dataset characteristics.

Table 4: Clinical constraint violation rates compared to TimeDiff Tian et al. (2023). **C1** (mortality monotonicity): the mortality flag, once raised, must remain set for all subsequent timesteps. **C2** (NaN-after-death): all measurement NaN-flags must be set after the first death timestep. Rates are computed among generated patients who ever die (N_{dead}). MIRRORTD enforces both constraints by construction; TimeDiff has no such mechanism.

Dataset	Model	N_{dead} / N_{synth}	C1 violation ↓	C2 violation ↓
MIMIC-III	TimeDiff	715 / 20,000	77.9%	27.8%
	MirrorTD	218 / 20,480	**4.1%**	**0.9%**
MIMIC-IV	TimeDiff	6,169 / 20,000	55.0%	95.3%
	MirrorTD	0 / 20,480	**0.0%**	**0.0%**

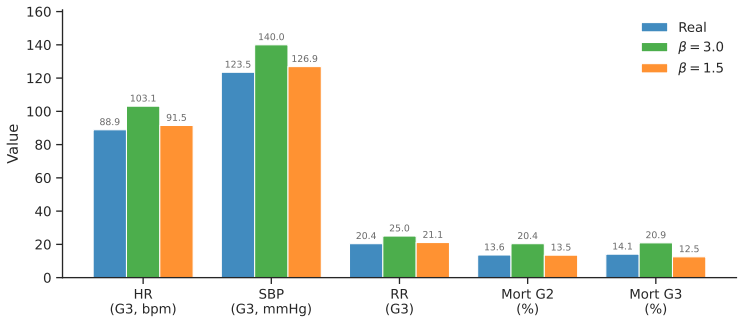

Figure 5: Effect of guidance scale β on Group 3 vital signs and Groups 2–3 mortality (Acuity model, MIMIC-III). $\beta = 1.5$ recovers the real distribution faithfully; $\beta = 3.0$ over-amplifies severity.

4.10 Conditional Generation via Classifier-Free Guidance

To evaluate the capability of classifier-free guidance of MIRRORTD, we use the model to generate data following two four-class conditioning schemes ($C = 4$).

Conditioning schemes. *Scheme 1 — Acuity (severity quartiles).* A scalar severity score s_i is computed for each patient as the mean absolute standardized deviation of observed vital signs from the training-population median:

$$s_i = \frac{1}{|\mathcal{O}_i|} \sum_{(k,t)\in\mathcal{O}_i} \frac{|x_{i,k,t} - \widetilde{\mu}_k|}{\widetilde{\sigma}_k}, \quad \mathcal{O}_i = \{(k,t) : f_{i,k,t} = 0\}, \tag{4.1}$$

where $f_{i,k,t} = 1$ denotes a missing observation. Patients are split into four equal-sized quartiles of s_i (Group 0 = least abnormal, Group 3 = most abnormal; ≈3,347 patients per group).

Scheme 2 — Mortality × Severity (MortSev). The score s_i is paired with the binary mortality outcome via a 2×2 split at the within-stratum median, yielding: Group 0 (survived, low s_i, N=6,154), Group 1 (survived, high s_i, N=6,169), Group 2 (died, low s_i, N=547), Group 3 (died, high s_i, N=518). The resulting 12:1 class imbalance is handled by `WeightedRandomSampler`, which up-weights each decedent sample so that all four classes contribute equally in expectation per batch.

Figure 5 compares guidance scales $\beta \in 1.5, 3.0$ against the real distribution for key metrics of the highest-acuity group. $\beta = 3.0$ over-amplifies Group 3 vital signs (HR +16%, SBP +13%) and inflates mortality rates, while $\beta = 1.5$ brings generated vital signs within 3% of true means and reproduces Group 2 mortality nearly the same (13.5% vs. 13.6%). So, we use $\beta = 1.5$ for subsequent experiments.

Acuity results ($\beta = 1.5$). Figure 6 shows per-group real and generated means for HR, RR, SBP, and mortality. The model captures the monotonic increase of RR and SBP from Group 0 to Group 3 and correctly assigns near-zero mortality to Groups 0–1 while generating ≈13 % mortality for Groups 2–3, matching the

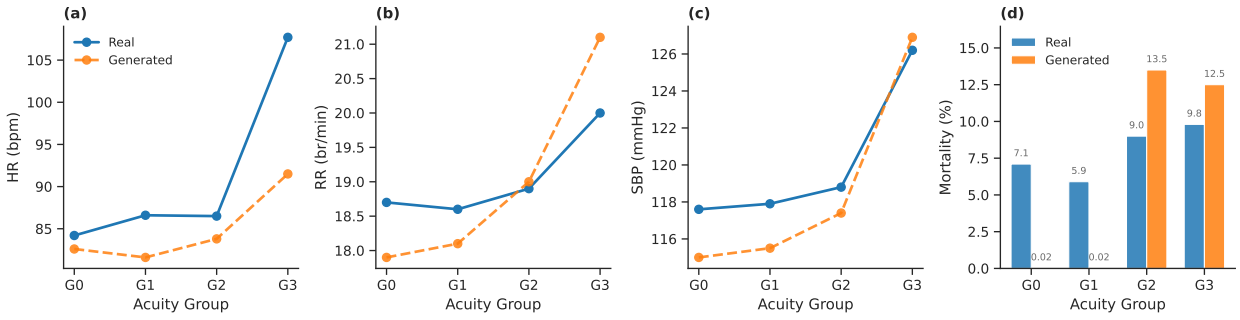

Figure 6: Acuity CFG results ($\beta = 1.5$, MIMIC-III, $N_{\text{gen}} = 10{,}240$ per group). **(a–c)** Real (solid) and generated (dashed) means for HR, RR, and SBP across four acuity groups. **(d)** Mortality rate per group; the model correctly assigns near-zero mortality to low-acuity groups and ≈13 % to high-acuity groups.

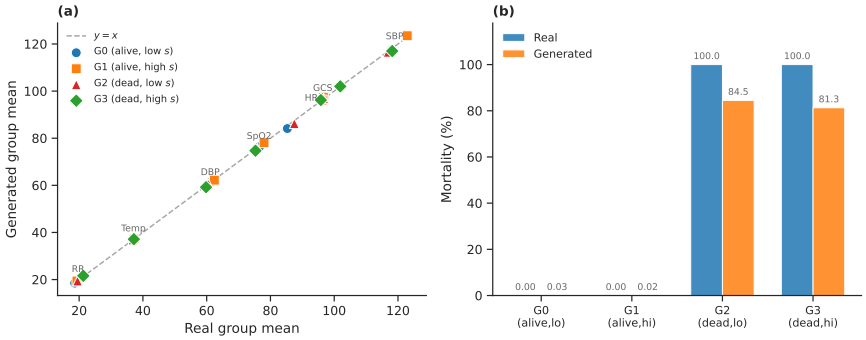

Figure 7: MortSev CFG results ($\beta = 1.5$, MIMIC-III). **(a)** Scatter of real vs. generated per-group vital-sign means for all seven channels × four groups (28 points total, coloured by group). All points lie near $y = x$, indicating faithful reproduction of within-group vital-sign distributions. **(b)** Mortality rate by group: survivors (G0–G1) are near zero; decedents (G2–G3) reach 81–85 %, compared to 100 % real.

true rates. HR separation is weaker for Groups 0–2 because the within-ICU acuity differences are naturally small; the large spike at Group 3 is partially recovered (91.5 vs. 107.7 bpm real).

MortSev results ($\beta = 1.5$). Figure 7 summarises the MortSev model. Panel (a) plots real vs. generated per-group vital-sign means across all seven channels; all 28 points lie close to the $y = x$ diagonal, indicating reproduction of within-group distributions within 1–2 %. Panel (b) shows clear separation between survivors (mortality ≈ 0 % for Groups 0–1) and decedents (mortality 81–85 % for Groups 2–3). The shortfall from 100 % in decedent groups reflects the 12:1 class imbalance during training, only partially corrected by `WeightedRandomSampler`. Compared with Acuity, MortSev yields sharper group separation due to the strong, unambiguous mortality signal used for conditioning.

5 Related Work

Generative modeling for EHRs has advanced rapidly due to growing interest in privacy-preserving data sharing and robust model development. Early approaches are dominated by GAN-based models (Choi et al., 2017; Baowaly et al., 2019; Esteban et al., 2017; Yoon et al., 2023), which struggle with instability and mode collapse, particularly for long and irregular clinical trajectories. VAE-based approaches (Biswal et al., 2021) and autoregressive models (Theodorou et al., 2023) improve stability but remain limited in capturing multimodal features and complex temporal dependencies.

Recently, diffusion models have gained popularity for both static (Kim et al., 2022; Lee et al., 2023; Kotelnikov et al., 2023; Zhang et al., 2023; He et al., 2023; Yuan et al., 2023a; Ceritli et al., 2023; Naseer et al., 2023; Han et al., 2024) and longitudinal EHR synthesis (Chen et al., 2021; Nicholas et al., 2023; Li et al., 2023; Tian et al., 2023; He et al., 2024; Deng et al., 2025; Cho et al., 2025; Ibrahim et al., 2025; Chadalawada & Bukaita, 2025). While TimeDiff (Tian et al., 2023) supports mixed-type EHR generation, it cannot enforce structural constraints such as mortality progression or NaN validity, which are essential for producing clinically meaningful sequences. In contrast, the proposed MirrorTD explicitly incorporates constraint-aware latent representations and guarantees validity by construction, bridging a gap left open by all prior EHR diffusion models.

Handling discrete time-series generation is challenging due to non-continuous state spaces and temporal dependencies in mixed-type EHR data. GAN-based architectures with RNNs (Esteban et al., 2017; Yoon et al., 2019; Jeon et al., 2022) have been proposed but struggle with long-horizon coherence and heterogeneous features. Diffusion-based approaches (Tashiro et al., 2021; Austin et al., 2021; Alcaraz & Strodthoff, 2022; Biloš et al., 2023; Yuan et al., 2023b) address some limitations via structured noise processes, while autoregressive Transformers (Salinas et al., 2020; Zhou et al., 2021; Wu et al., 2021; Li et al., 2022) improve long-range forecasting but remain sequential. However, these methods focus on homogeneous categorical sequences and do not model joint numerical, categorical, and constraint-bearing features. MirrorTD extends diffusion-based modeling by integrating Gaussian and multinomial diffusion in a unified architecture for mixed-type EHR signals.

Incorporating domain constraints into generative models is essential for realistic and safe synthetic data. Traditional approaches include constrained VAEs (Kingma et al., 2018), rule-based filtering, or post-processing, which risk infeasible sequences or distort the learned distribution. More recent methods integrate constraints directly via latent relaxations or convexified sets (Liu et al., 2023; Feng et al., 2024), improving guarantees while maintaining expressiveness. Liu et al. (2023) proposed mirror diffusion with closed-form mappings for convex sets, e.g., ℓ_2 ball and simplices. Feng et al. (2024) extended these mappings to neural networks (MLPs). MirrorTD builds on this by introducing a mirror-mapping VAE for temporally structured mortality and NaN constraints within a spatio-temporal transformer.

6 Conclusion

We introduced MirrorTD, a unified score-based generative framework for mixed-type EHR time series that jointly addresses heterogeneous feature modeling, strict clinical constraints, and flexible conditional synthesis. By integrating Gaussian and discrete diffusion processes with a mirror-mapping VAE, MirrorTD embeds mortality and missingness constraints into a continuous latent space, substantially reducing clinical-constraint violations relative to unconstrained baselines. The proposed spatio-temporal attention architecture captures long-range temporal patterns and complex cross-feature dependencies, and classifier-free guidance enables controllable conditional generation for rare patient subpopulations.

Across three ICU benchmarks (MIMIC-III, MIMIC-IV, eICU) and five baselines, MirrorTD achieves the lowest mean discriminative score on every dataset—most notably on eICU (0.032 vs. 0.263 for TimeDiff) and MIMIC-IV (0.161 vs. 0.245)—the closest Markov transition fidelity (0.227 vs. 0.430 on MIMIC-III), and reduces hard-constraint violations by over an order of magnitude (4.1% vs. 77.9% for mortality monotonicity on MIMIC-III, with residual violations attributable to binarisation rounding), while remaining competitive on downstream mortality prediction and privacy.

Limitations and future work. Our framework assumes a fixed, regular observation grid; extending it to irregularly sampled time series—common in outpatient EHRs—would require coupling the diffusion process with a continuous-time model such as a neural ODE. The evaluation is also restricted to ICU cohorts from three US-based databases; generalization to other clinical domains (e.g., longitudinal outpatient records) and non-US populations remains to be validated. Additionally, the mirror-mapping VAE currently handles only the two structural constraints formalized in Section 2; incorporating richer domain knowledge (e.g., physiological bounds or pharmacokinetic consistency) is a natural next step. Finally, integrating MirrorTD with RL environments for treatment-policy simulation is a promising direction for future work.

References

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Mohammad Kamrul Hasan Baowaly, Chia Hung Lin, Chih-Wei Liu, and Chih-Long Chen. Synthesizing electronic health records using improved generative adversarial networks. In *Journal of the American Medical Informatics Association*, volume 26, pp. 228–241. Oxford University Press, 2019.

Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion. In *International Conference on Machine Learning*, pp. 2452–2470. PMLR, 2023.

Siddharth Biswal, Soumya Ghosh, Jon Duke, Bradley Malin, Walter Stewart, Cao Xiao, and Jimeng Sun. Eva: Generating longitudinal electronic health records using conditional variational autoencoders. In *Machine Learning for Healthcare Conference*, pp. 260–282. PMLR, 2021.

Taha Ceritli, Ghadeer O Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P Creagh, and David A Clifton. Synthesizing mixed-type electronic health records using diffusion models. *arXiv preprint arXiv:2302.14679*, 2023.

Priyatham Chadalawada and Wisam Bukaita. Balancing privacy and data utility in electronic health records: A two-stage synthetic data generation approach. *Medical Research Archives*, 13(10), 2025.

Raymond Chen, Yuxin Luo, Edward Choi, and David Sontag. Synthetic electronic health records generated with variational graph autoencoders. *Scientific Reports*, 11(1):1–12, 2021.

Eunbyeol Cho, Jiyoun Kim, Minjae Lee, Sungjin Park, and Edward Choi. Generating multi-table time series ehr from latent space with minimal preprocessing. *arXiv preprint arXiv:2507.06996*, 2025.

Edward Choi, Sushant Biswal, Bradley Malin, Jon Duke, Walter Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference*, pp. 286–305. PMLR, 2017.

Bowen Deng, Chang Xu, Hao Li, Yu-hao Huang, Min Hou, and Jiang Bian. Tardiff: Target-oriented diffusion guidance for synthetic electronic health record time series generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 474–485, 2025.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar R"atsch. Real-valued (medical) time series generation with recurrent conditional gans. In *Advances in neural information processing systems*, pp. 4628–4638, 2017.

Berthy T Feng, Ricardo Baptista, and Katherine L Bouman. Neural approximate mirror maps for constrained diffusion models. *arXiv preprint arXiv:2406.12816*, 2024.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101 (23):e215–e220, 2000.

Jun Han, Zixiang Chen, Yongqian Li, Yiwen Kou, Eran Halperin, Robert E Tillman, and Quanquan Gu. Guided discrete diffusion for electronic health record generation. *arXiv preprint arXiv:2404.12314*, 2024.

Huan He, Shifan Zhao, Yuanzhe Xi, and Joyce C Ho. Meddiff: Generating electronic health records using accelerated denoising diffusion model. *arXiv preprint arXiv:2302.04355*, 2023.

Huan He, William Hao, Yuanzhe Xi, Yong Chen, Bradley Malin, and Joyce Ho. A flexible generative model for heterogeneous tabular {EHR} with missing modality. In *The Twelfth International Conference on Learning Representations*. https://openreview. net/forum? id= W2tCmRrj7H, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

Mahmoud Ibrahim, Bart Elen, Chang Sun, Gökhan Ertaylan, and Michel Dumontier. Enabling granular subgroup level model evaluations by generating synthetic medical time series. *arXiv preprint arXiv:2510.19728*, 2025.

Jinsung Jeon, Jeonghak Kim, Haryong Song, Seunghyeon Cho, and Noseong Park. Gt-gan: General purpose time series synthesis with generative adversarial networks. *Advances in Neural Information Processing Systems*, 35:36999–37010, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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.

Jayoung Kim, Chaejeong Lee, and Noseong Park. Stasy: Score-based tabular data synthesis. *arXiv preprint arXiv:2210.04018*, 2022.

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Variational autoencoders with arbitrary conditioning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2363–2372. PMLR, 2018.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pp. 18940–18956. PMLR, 2023.

Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1):1–12, 2023.

X Li et al. Transformer-based models for generating synthetic time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7712–7720, 2022.

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36:42898–42917, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

Ahmed Ammar Naseer, Benjamin Walker, Christopher Landon, Andrew Ambrosy, Marat Fudim, Nicholas Wysham, Botros Toro, Sumanth Swaminathan, and Terry Lyons. Scoehr: generating synthetic electronic health records using continuous-time diffusion models. In *Machine Learning for Healthcare Conference*, pp. 489–508. PMLR, 2023.

I Nicholas, Hsien Kuo, Federico Garcia, Anders Sonnerborg, Michael Bohm, Rolf Kaiser, Maurizio Zazzi, Louisa Jorm, and Sebastiano Barbieri. Synthetic health-related longitudinal data with mixed-type variables generated using diffusion models. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

Giannis Nikolentzos, Michalis Vazirgiannis, Christos Xypolopoulos, Markus Lingman, and Erik G Brandt. Synthetic electronic health records generated with variational graph autoencoders. *NPJ Digital Medicine*, 6(1):83, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Noam Hajaj, Michael Hardt, Peter J Liu, Xing Liu, Joseph Marcus, Mia Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):1–10, 2018.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34: 24804–24816, 2021.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, 14(1):5305, 2023.

Muhang Tian, Bernie Chen, Allan Guo, Shiyi Jiang, and Anru R Zhang. Reliable generation of privacy-preserving synthetic ehr time series via diffusion models. *arXiv preprint arXiv:2310.15290*, 2023.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430, 2021.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020. doi: 10.1016/j.neucom.2019.12.136.

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1):7609, 2022. doi: 10.1038/s41467-022-35295-1.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in neural information processing systems*, pp. 5508–5518, 2019.

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ digital medicine*, 6(1):141, 2023.

Hongyi Yuan, Songchi Zhou, and Sheng Yu. Ehrdiff: Exploring realistic ehr synthesis with diffusion models. *arXiv preprint arXiv:2303.05656*, 2023a.

Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3173–3184, 2023b.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

A Implementation Details

A.1 Datasets and Data Preprocessing

W preprocess the datasets following Tian et al. (2023), with the only difference being the time-dependent mortality flag. We extract the mortality time from the raw dataset, and, instead of appending a column of all-zeros or all-ones, we append a column made up of time-dependent mortality flags. We show the results of the precessing in Table 5.

Table 5: Basic information of the MIMIC-III, MIMIC-IV, and eICU datasets after preprocessing. We display the number of datapoints used in training, the length of time series, and the number of features belonging to each type.

Dataset	Size	L	d_{num}	d_{cat}	d_{nan}	Has $\mathbf{x}_{\text{mort}}$?
MIMIC-III	20918	25	7	0	7	Yes
MIMIC-IV	20400	72	8	3	8	Yes
eICU	91356	276	4	0	4	Yes

A.2 Diffusion Model

For both the Gaussian and discrete diffusion components, let $\{\beta^{(t)}\}_{t\in[T]}$ denote the noise schedule, and define $\alpha^{(t)} = 1 - \beta^{(t)}$ and $\bar{\alpha}^{(t)} = \prod_{s=1}^{t} \alpha^{(s)}$ for all $t \in [T]$ (Ho et al., 2020). We now describe each component separately. In the experiments, we use 50 diffusion steps in the diffusion model. Following Tian et al. (2023), we use the cosine noise schedule β_t.

Gaussian diffusion model. We utilize Gaussian diffusion for the combination of numerical features and the flag embeddings (See Section 3.2 for details) $\mathbf{x}_{\text{g}} = [\mathbf{x}_{\text{num}}, \mathbf{z}_{\text{flag}}]$. In the Gaussian diffusion model, the conditional distribution $q(\mathbf{x}_{\text{g}}^{\cdot,l,(t)} | \mathbf{x}_{\text{g}}^{\cdot,l,(t-1)})$ is given by:

$$\mathbf{x}_{\text{g}}^{\cdot,l,(t)} | \mathbf{x}_{\text{g}}^{\cdot,l,(t-1)} \sim \mathcal{N}(\sqrt{1-\beta^{(t)}} \cdot \mathbf{x}_{\text{g}}^{\cdot,l,(t-1)}, \beta^{(t)} \cdot \mathbf{I}).$$

Therefore, given the clean data $\mathbf{x}_{\cdot,l}^{\text{g},(0)}$, we can sample the noisy data $\mathbf{x}_{\cdot,l}^{\text{g},(t)}$ as

$$\mathbf{x}_{\cdot,l}^{\text{g},(t)} = \sqrt{\bar{\alpha}^{(t)}} \cdot \mathbf{x}_{\cdot,l}^{\text{g},(0)} + \sqrt{1-\bar{\alpha}^{(t)}} \cdot \boldsymbol{\epsilon}, \tag{A.1}$$

where $\boldsymbol{\epsilon}$ is the standard gaussian noise. We use the neural network $\mathbf{s}_{\mathrm{g}}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)})$ to predict the noise $\boldsymbol{\epsilon}$. During inference, we sample $\mathbf{x}_{\cdot,l}^{\mathrm{g},(T)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and for $t = T, T-1, \ldots, 1$, we iteratively sample

$$\mathbf{x}_{\mathrm{g}}^{\cdot,l,(t-1)} | \mathbf{x}_{\mathrm{g}}^{\cdot,l,(t)} \sim \mathcal{N}\Big(\boldsymbol{\mu}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}), \frac{1-\bar{\alpha}^{(t-1)}}{1-\bar{\alpha}^{(t)}}\beta^{(t)} \cdot \mathbf{I}\Big), \text{ where}$$

$$\boldsymbol{\mu}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)}) = \frac{1}{\sqrt{\alpha^{(t)}}}\Big(\mathbf{x}_{\mathrm{g}}^{\cdot,l,(t)} - \frac{\beta^{(t)}}{\sqrt{1-\bar{\alpha}^{(t)}}} \cdot \mathbf{s}_{\mathrm{g}}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)})\Big).$$

The ELBO training objective of the Gaussian diffusion model is

$$\mathcal{L}^{\mathrm{g}}(\boldsymbol{\theta}) = \mathbb{E}_t \mathbb{E}_{\mathbf{x}^{(0)};\boldsymbol{\epsilon}} \frac{\beta^{(t)}}{\alpha^{(t)}(1-\bar{\alpha}^{(t-1)})} \|\boldsymbol{\epsilon} - \mathbf{s}^{t,\boldsymbol{\theta}}(\mathbf{x}^{(t)})\|_2^2.$$

Discrete diffusion model. We use the multinomial diffusion model (Hoogeboom et al., 2021) for the categorical features $\mathbf{x}_{\mathrm{cat}}$. Let $\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} \in \{0,1\}^K$ also denote the one-hot representation, where K is the number of classes. The conditional distribution $q(\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)})$ is given by:

$$\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} \sim \mathcal{C}((1-\beta^{(t)}) \cdot \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} + \beta^{(t)}/K).$$

Given the clean data $\mathbf{x}_{\cdot,l}^{\mathrm{cat},(0)}$, we can also sample the noisy categorical features as:

$$\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(0)} \sim \mathcal{C}(\bar{\alpha}^{(t)} \cdot \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(0)} + (1-\bar{\alpha}^{(t)})/K). \tag{A.2}$$

In the reverse process, the conditional distribution $\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)}$ satisfies

$$p(\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)}) = \sum\nolimits_{\hat{\mathbf{x}}_{\mathrm{cat}}^{\cdot,l,(0)}} p(\hat{\mathbf{x}}_{\mathrm{cat}}^{\cdot,l,(0)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)}) \cdot q(\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} | \hat{\mathbf{x}}_{\mathrm{cat}}^{\cdot,l,(0)}, \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)}). \tag{A.3}$$

We train the score network $\mathbf{s}_{\mathrm{cat}}^{\boldsymbol{\theta},t}(\mathbf{x}^{(t)})$ to approximate the logits of $p(\hat{\mathbf{x}}_{\mathrm{cat}}^{\cdot,l,(0)} | \mathbf{x}^{(t)})$. Finally, we calculate $p^{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)})$ following (A.3) combined with the property

$$\mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t-1)} | \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(0)}, \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} \sim \mathcal{C}(\boldsymbol{\phi}/(\mathbf{1}^\top \boldsymbol{\phi})),$$
$$\boldsymbol{\phi} = \big((1-\beta^{(t)}) \cdot \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(t)} + \beta^{(t)}/K\big) \odot \big(\bar{\alpha}^{(t-1)} \mathbf{x}_{\mathrm{cat}}^{\cdot,l,(0)} + (1-\bar{\alpha}^{(t)})/K\big).$$

The ELBO training objective of the multinomial diffusion model is

$$\mathcal{L}^{\mathrm{cat}}(\boldsymbol{\theta}) = \mathbb{E}_t \mathbb{E}_{\mathbf{x}^{(0)};\mathbf{x}^{(t)}} \big[D_{\mathrm{KL}}(q(\mathbf{x}_{\mathrm{cat}}^{(t-1)} | \mathbf{x}^{(t)}, \mathbf{x}^{(0)}) || p^{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{cat}}^{(t-1)} | \mathbf{x}^{(t)}))\big].$$

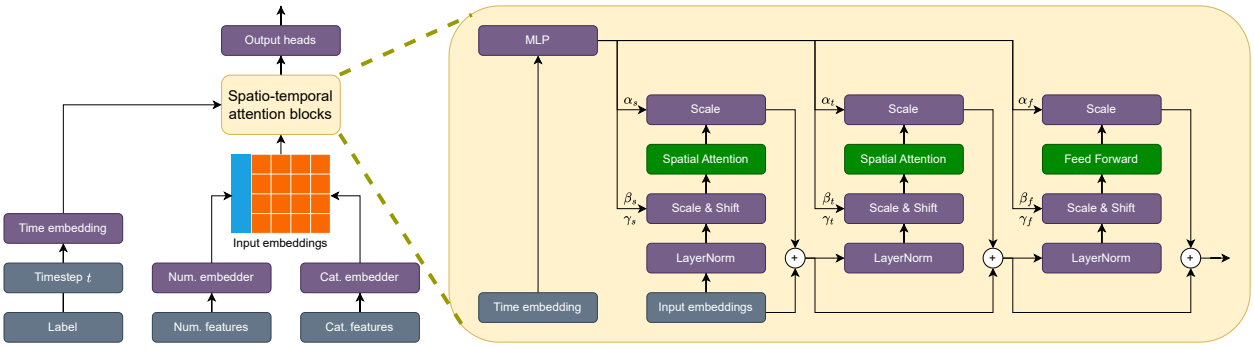

Figure 8: Spatio-temporal transformer with AdaLN.

Score Network. We use adaptive layernorm (AdaLN) following Peebles & Xie (2023). Different from the original AdaLN that only deals with one attention block and one feed-forward block per layer, our model has two attention blocks and the feed-forward block, which requires an additional set of rescaling parameters. An illustration of our model is shown in Figure 8. We use 10 spatio-temporal attention blocks, and the embedding dimension is 256 for MIMIC-III and eICU, and 512 for MIMIC-IV.

A.3 Mirror Mapping

We use similar spatio-temporal transformer structures in the mirror mapping as in the score network. The only structural difference is that AdaLN as well as the time embedding is not used, and we perform the vanilla residual connection without rescaling.

During training, we use $d = 8$ for the flag embeddings of each flag feature, with a total of 4 attention layers. To ensure the abundance of training data with nontrivial mortality flags, we randomly perturb the raw data by randomly selecting the time of mortality of the data and setting all subsequent flags to 1.

Since the output of the decoder is in the relaxed set C^+, we round the output to $\{0, 1\}$ to project the output to the desired space of flags C.

A.4 Model Training

We use the Adam optimizer with learning rate 8e-5 and momentum parameters $(\beta_1, \beta_2) = (0.9, 0.99)$. The models are trained for 40000 steps.

B Additional Results

B.1 Prevalence

We now present the additional experiment results of the prevalence on MIMIC-III/IV and the eICU dataset. For clarity, we only present the results from MirrorTD, TimeDiff, and GT-GAN because the prevalance of DSPD/CSPD are out of the reasonable range. In terms of prevalence, MirrorTD performs in par with or better than TimeDiff, and significantly better than GT-GAN.

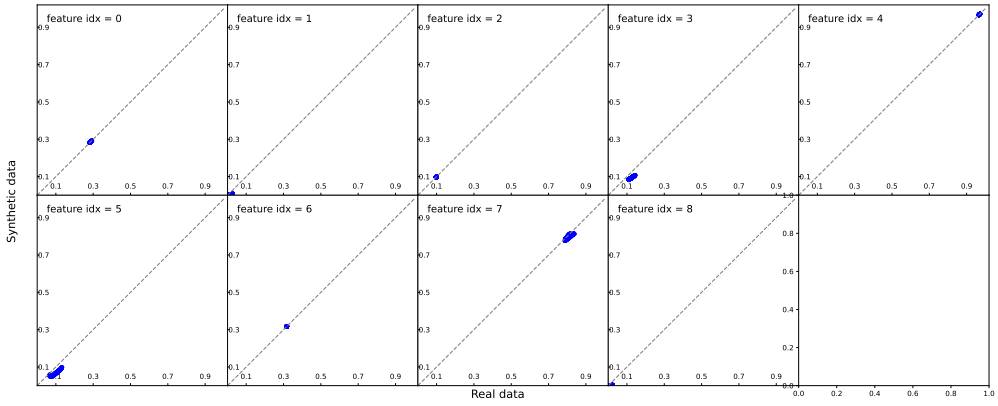

Figure 9: Prevalence of MirrorTD on the eICU dataset.

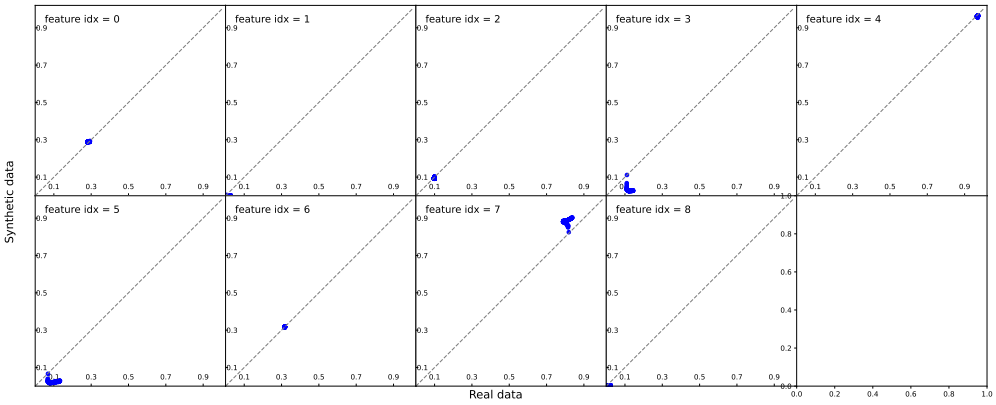

Figure 10: Prevalence of TimeDiff on the eICU dataset.

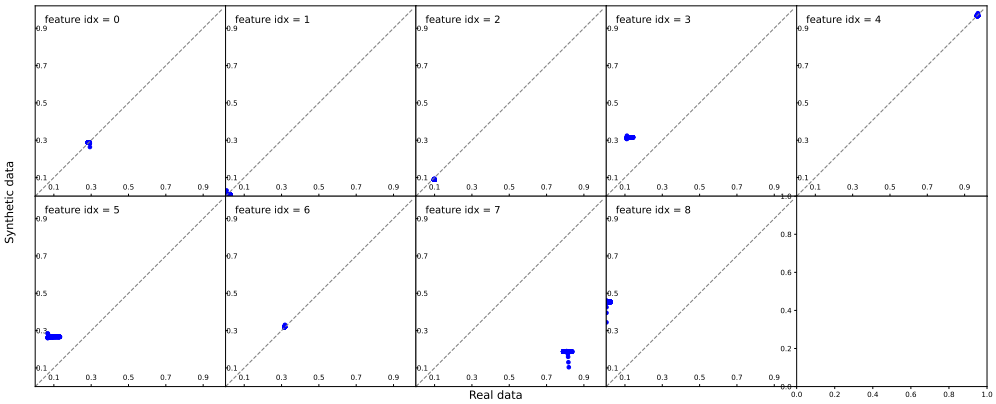

Figure 11: Prevalence of GT-GAN on the eICU dataset.

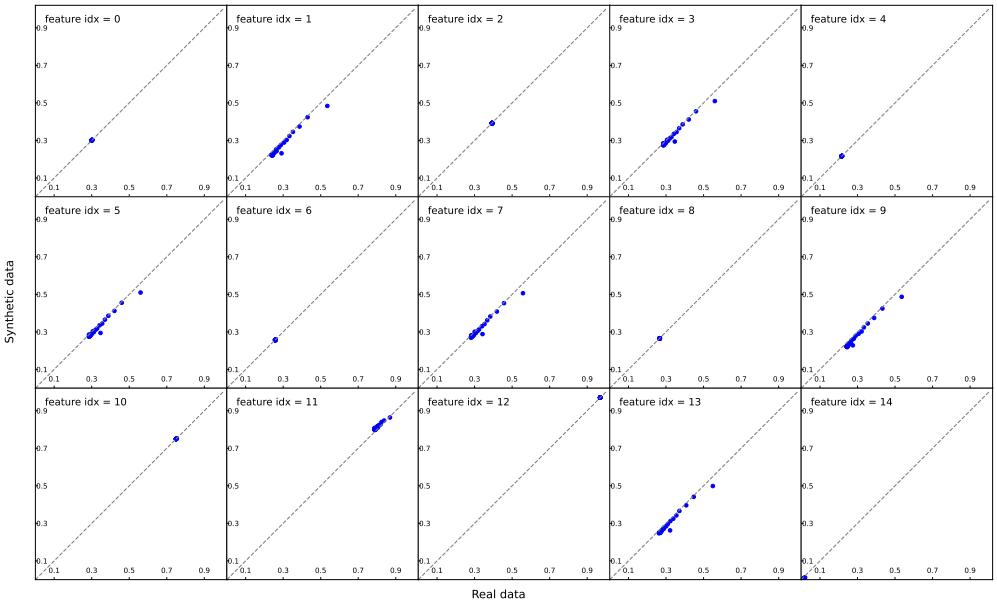

Figure 12: Prevalence of MirrorTD on the MIMIC-III dataset.

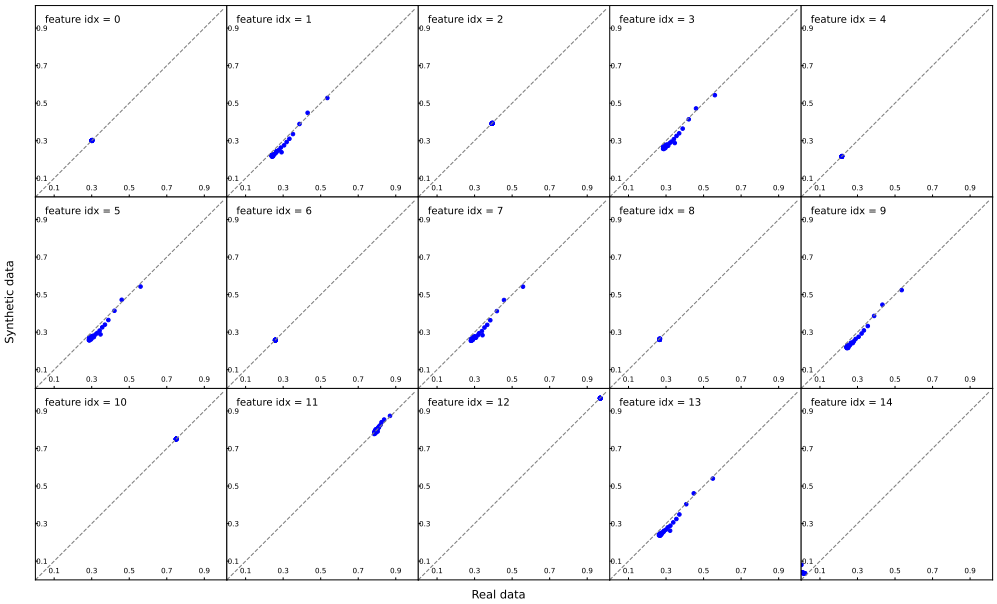

Figure 13: Prevalence of TimeDiff on the MIMIC-III dataset.

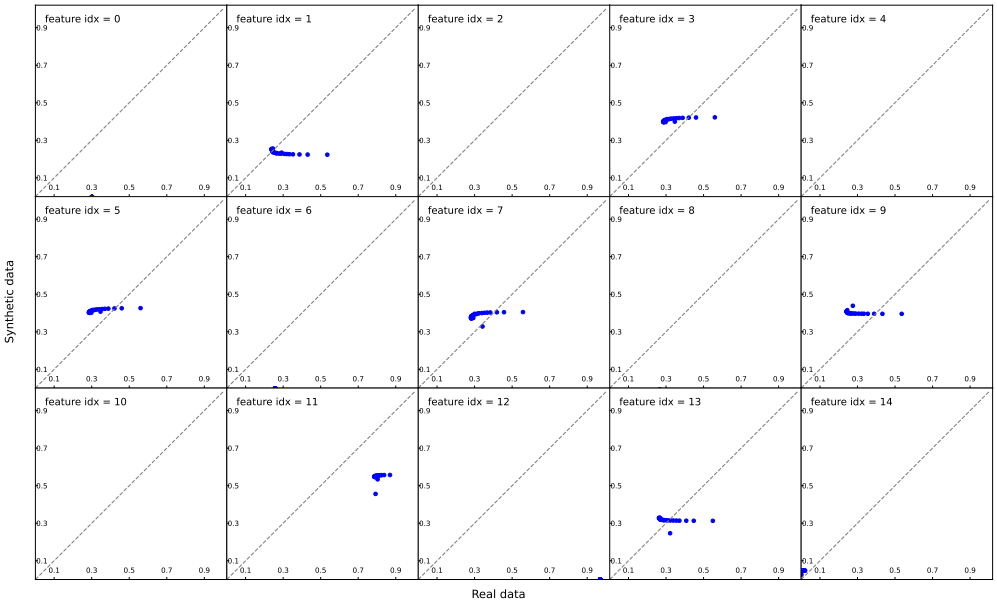

Figure 14: Prevalence of GT-GAN on the MIMIC-III dataset.

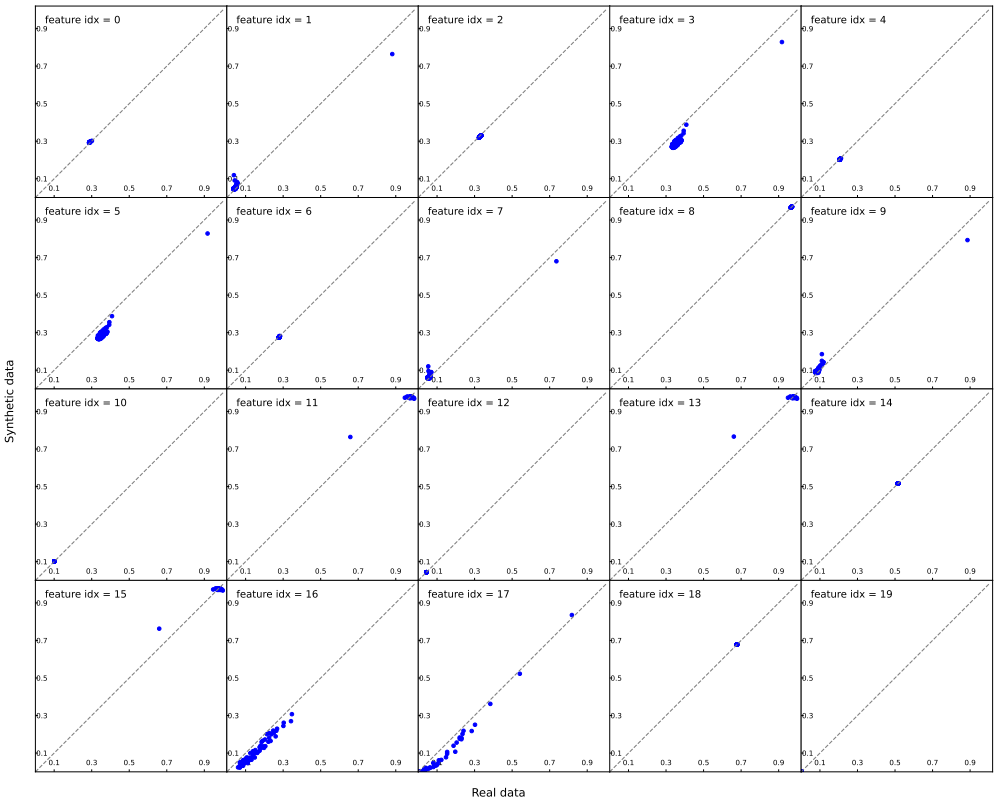

Figure 15: Prevalence of MirrorTD on the MIMIC-IV dataset.

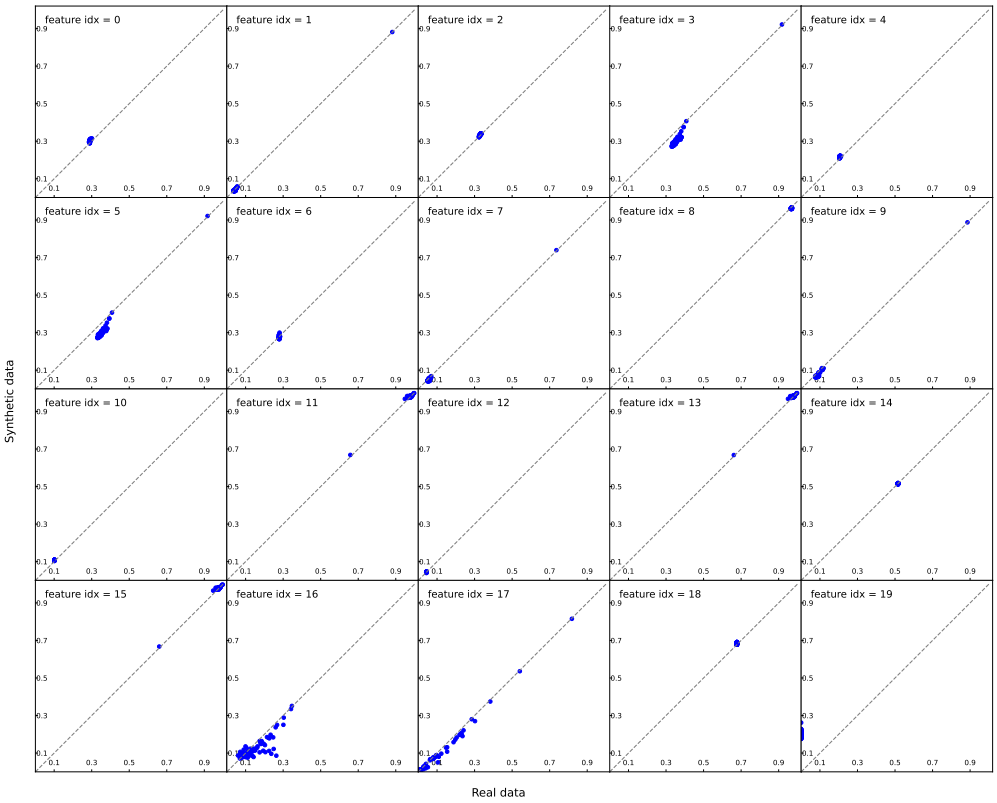

Figure 16: Prevalence of TimeDiff on the MIMIC-IV dataset.

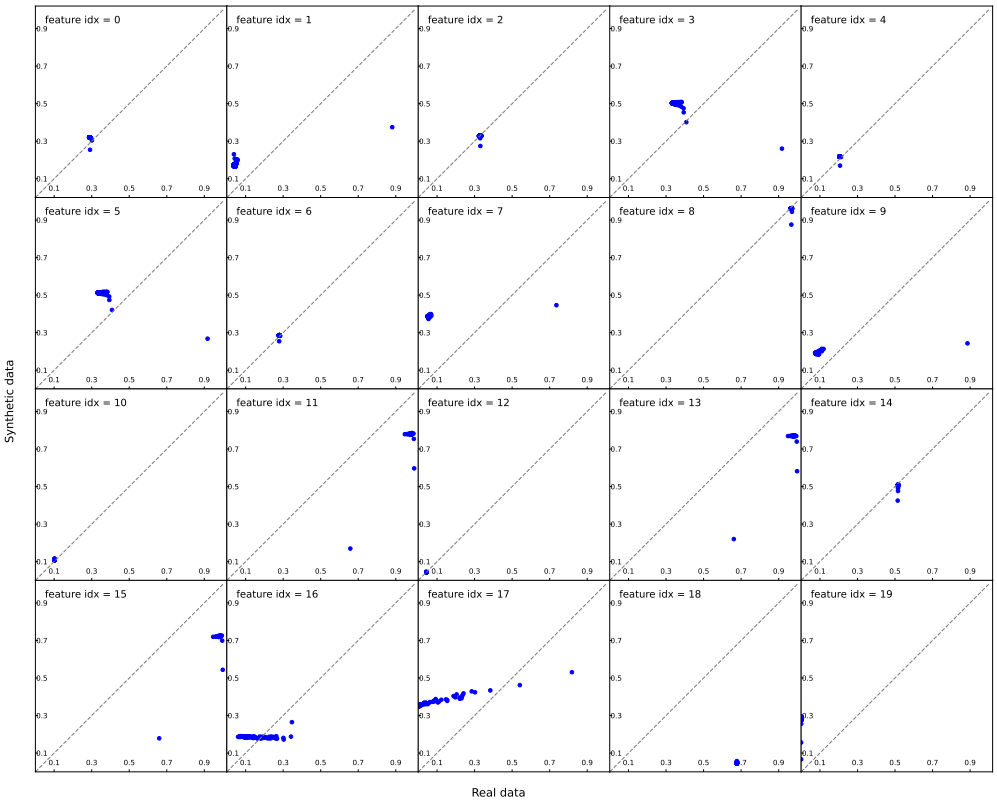

Figure 17: Prevalence of GT-GAN on the MIMIC-IV dataset.

B.2 MSM Transition Matrices

Now we show the experiment results around the MSM transition matrix. MirrorTD performs better than all other baselines on all three datasets, with significantly fewer entries of large error in the MSM transition matrix.

Table 6: MSM transition matrix distances of MirrorTD and baselines. The distances in settings where the MSM transition matrices cannot be calculated are reported as NaN.

Method	MIMIC-III	MIMIC-IV	eICU
MirrorTD	0.2274	0.2960	0.1706
TimeDiff	0.4297	0.5317	1.7193
DSPD-GP	1.3080	1.7735	NaN
DSPD-OU	2.0710	NaN	NaN
CSPD-GP	1.5027	1.6813	0.6819
CSPD-OU	1.4838	1.0938	NaN
GT-GAN	2.8172	2.5348	0.7847

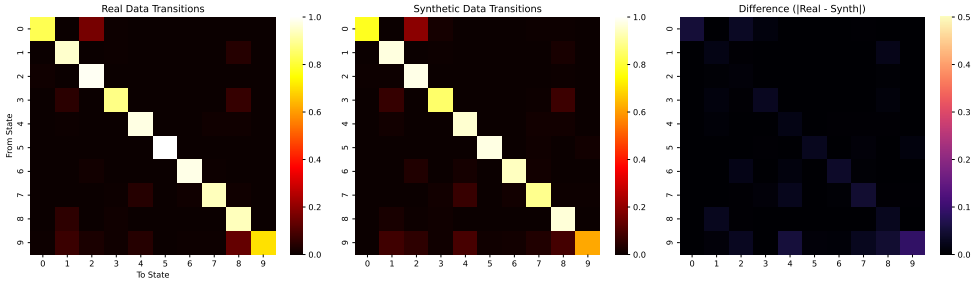

Figure 18: MSM transition matrix and error of MirrorTD on the eICU dataset.

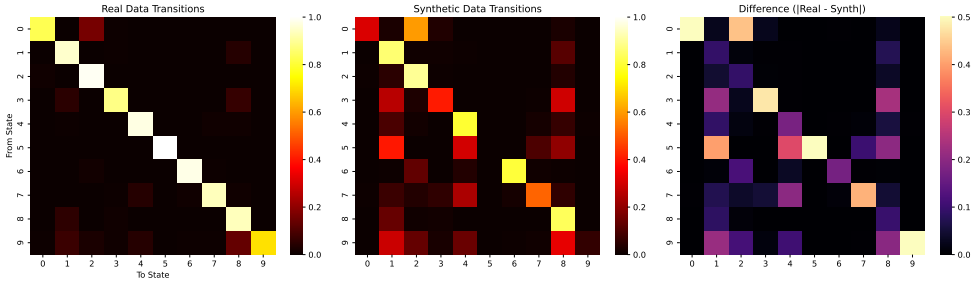

Figure 19: MSM transition matrix and error of TimeDiff on the eICU dataset.

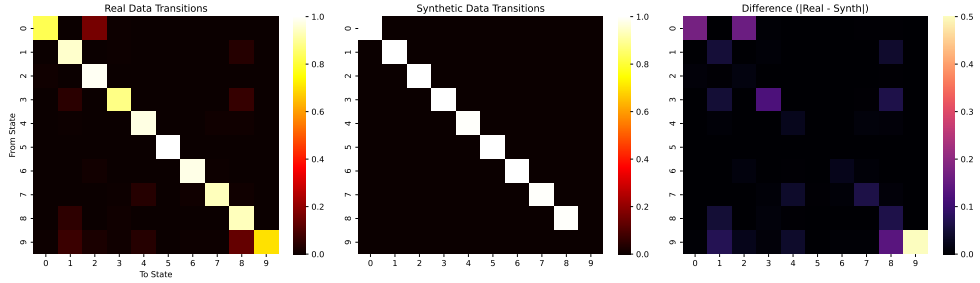

Figure 20: MSM transition matrix and error of GT-GAN on the eICU dataset.

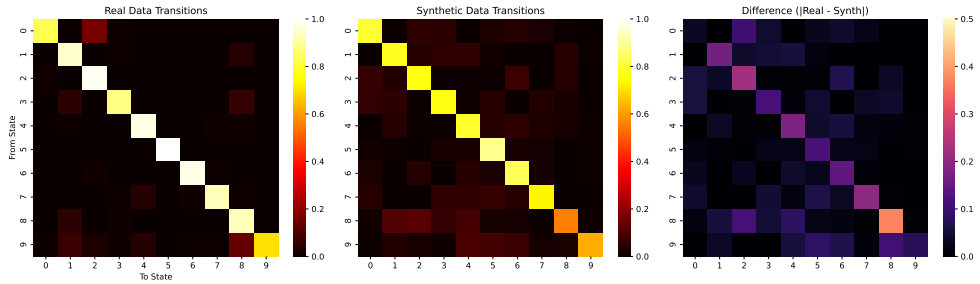

Figure 21: MSM transition matrix and error of CSPD-GP on the eICU dataset.

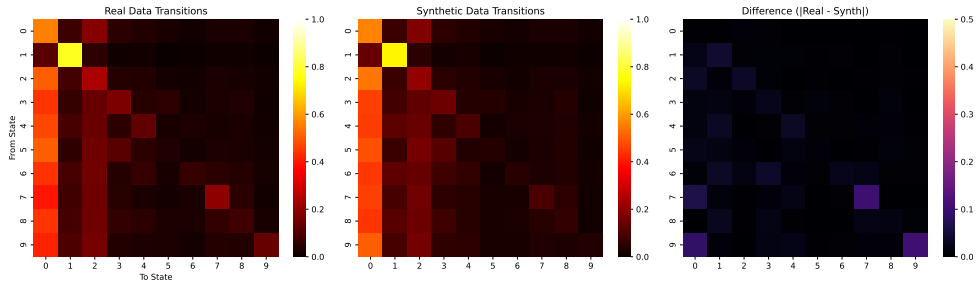

Figure 22: MSM transition matrix and error of MirrorTD on the MIMIC-III dataset.

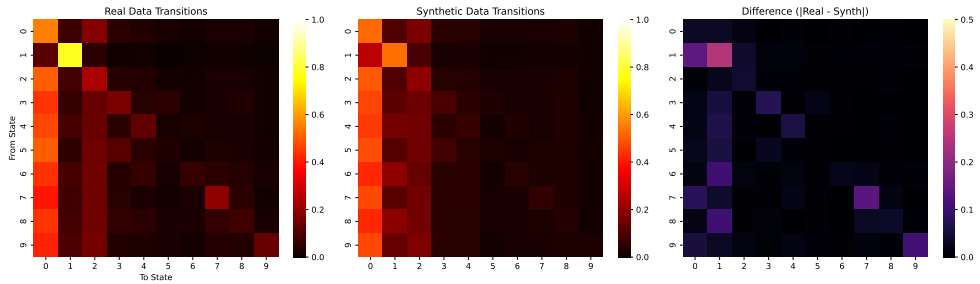

Figure 23: MSM transition matrix and error of TimeDiff on the MIMIC-III dataset.

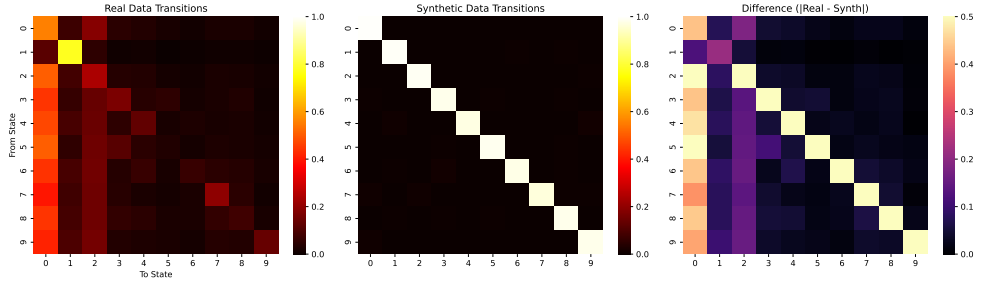

Figure 24: MSM transition matrix and error of GT-GAN on the MIMIC-III dataset.

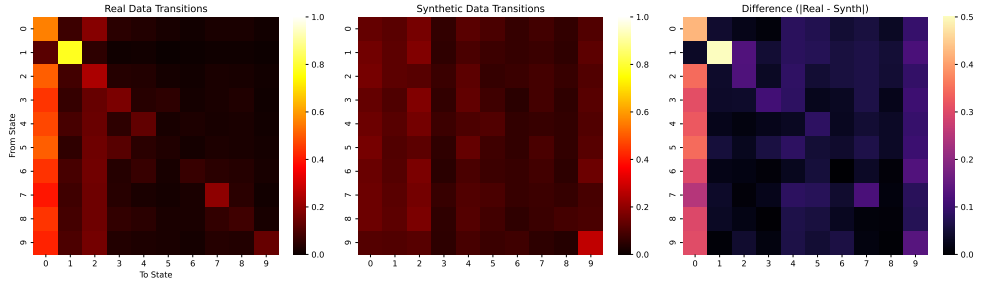

Figure 25: MSM transition matrix and error of DSPD-GP on the MIMIC-III dataset.

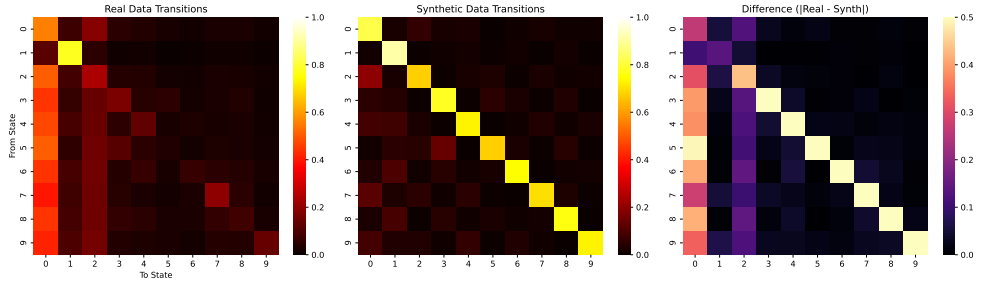

Figure 26: MSM transition matrix and error of CSPD-OU on the MIMIC-III dataset.

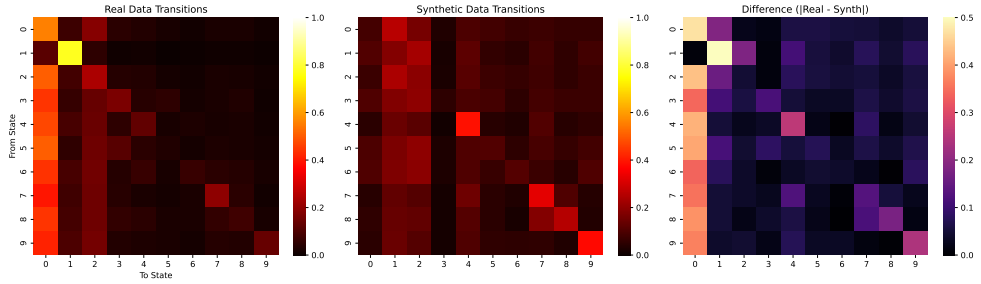

Figure 27: MSM transition matrix and error of CSPD-GP on the MIMIC-III dataset.

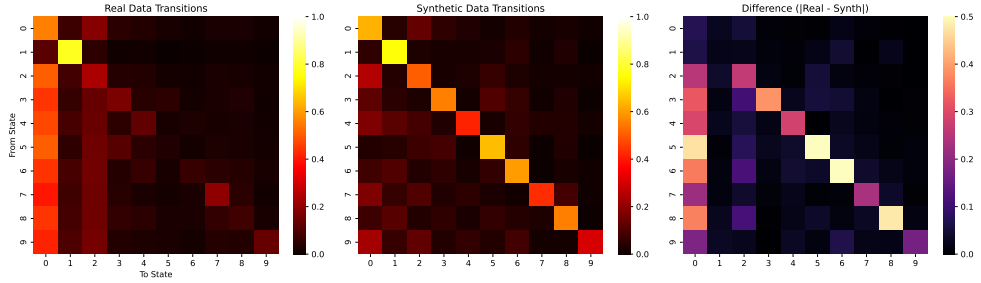

Figure 28: MSM transition matrix and error of CSPD-OU on the MIMIC-III dataset.

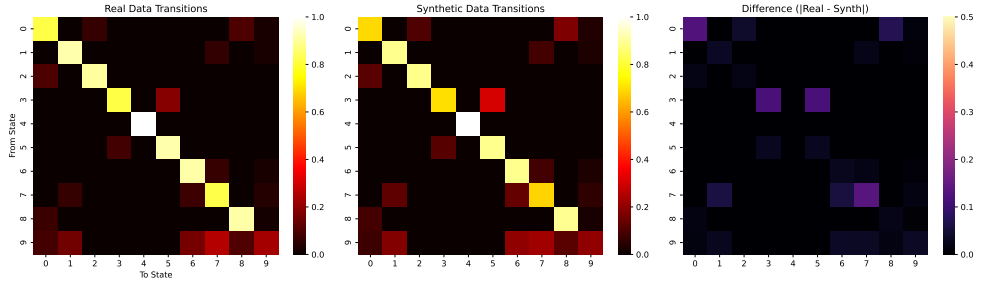

Figure 29: MSM transition matrix and error of MirrorTD on the MIMIC-IV dataset.

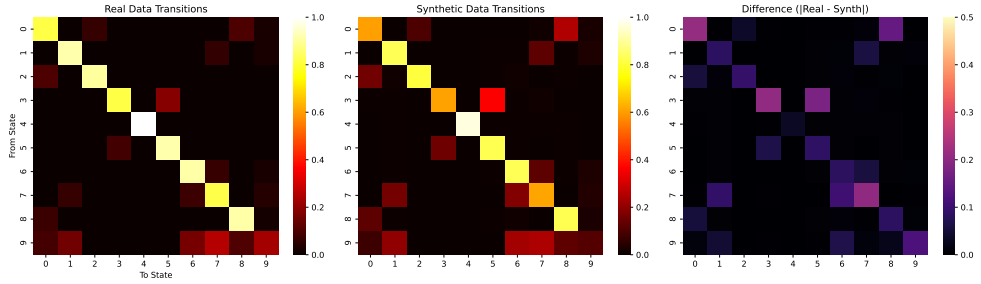

Figure 30: MSM transition matrix and error of TimeDiff on the MIMIC-IV dataset.

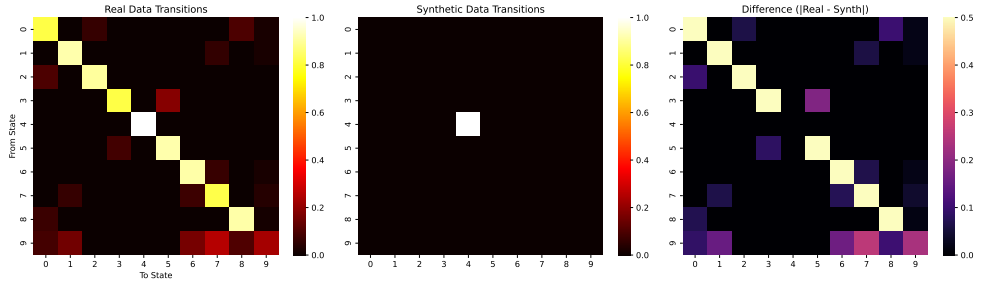

Figure 31: MSM transition matrix and error of GT-GAN on the MIMIC-IV dataset.

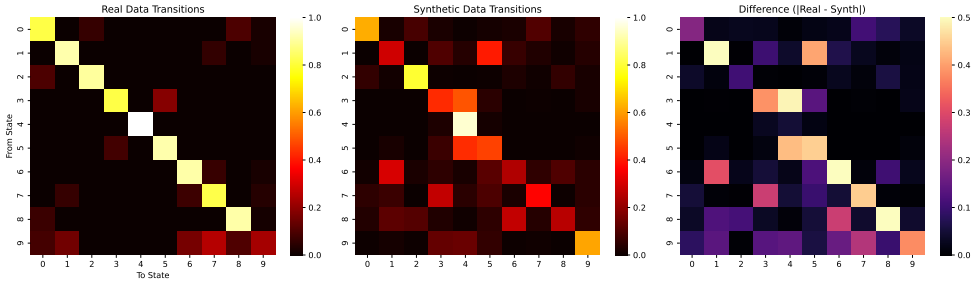

Figure 32: MSM transition matrix and error of DSPD-GP on the MIMIC-IV dataset.

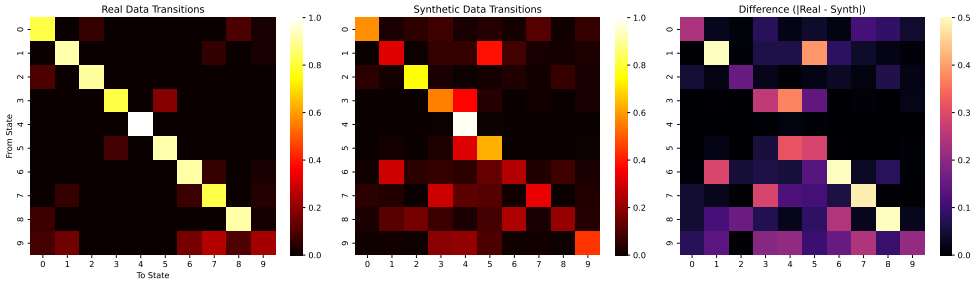

Figure 33: MSM transition matrix and error of CSPD-GP on the MIMIC-IV dataset.

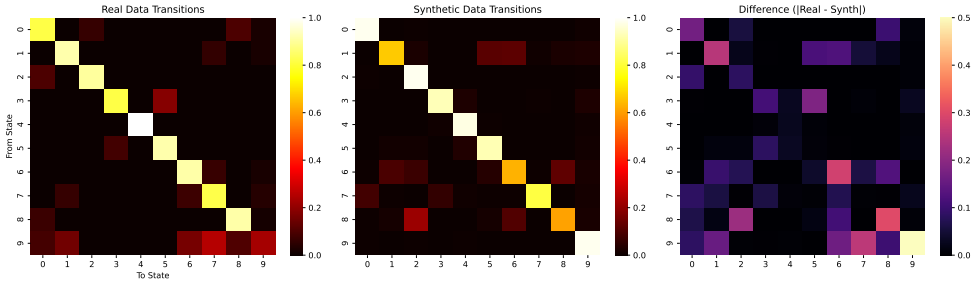

Figure 34: MSM transition matrix and error of CSPD-OU on the MIMIC-IV dataset.