
HealthForge: Advancing EHR Synthesis Using CFG-Based Conditional Diffusion in Latent Space

William Liu [†]

*Faculty of Arts and Science
University of Toronto*

williamc.liu@mail.utoronto.ca

Ryan Zhang [†]

*Faculty of Arts and Science
University of Toronto*

ryanz.zhang@mail.utoronto.ca

Taiyi Jin [†]

*Faculty of Arts and Science
University of Toronto*

tiger.jin@mail.utoronto.ca

Abdus Shaikh [†]

*Faculty of Arts and Science
University of Toronto*

abdus.shaikh@mail.utoronto.ca

Abstract

Synthetic Electronic Health Records (EHRs) provide a promising solution to the challenges of data scarcity and privacy constraints in clinical machine learning research. Diffusion models have recently emerged as strong generative approaches for high-dimensional health data, and recent existing systems such as EHRDiff (Wang et al., 2023) have proven the open-ended potential of diffusion models in this space. With proven potential, new applications emerge such as incorporating flexible conditional information in the generation. In this work, we introduce HealthForge, a diffusion-based framework built from scratch to support text-conditioned synthetic EHR generation.

Using the de-identified MIMIC-IV dataset, we first preprocess structured EHR vectors into a consistent training representation. To enable conditional generation, we employ large language models to derive concise clinical descriptions for each real patient record, and apply a secondary LLM-based filtering step to ensure that these labels are accurate and usable. The resulting text labels are embedded and provided as conditioning inputs to our denoising diffusion model, which is trained with classifier-free guidance (Ho & Salimans, 2022). To evaluate fidelity and privacy, we adopt the core utility (dimension-wise distribution alignment,

[†] Contributed equally. Code released in <https://github.com/UofT-CSC490-F2025/health-forge>

correlation structure) and privacy (attribute inference, membership inference) metrics established in prior EHR generation research, using them as standardized baselines to quantify improvements over existing methods. Our results indicate that HealthForge produces high-utility synthetic EHR data with generally strong, yet improvable, privacy behavior, while supporting controllable generation through natural-language conditioning.

1 Introduction and Problem Statement

Healthcare has become increasingly data-driven, yet access to real electronic health records (EHRs) remains tightly restricted due to privacy regulations, slow approval pipelines, and institutional barriers. These limitations make it difficult for many research groups to prototype machine-learning ideas or evaluate new modeling approaches using real clinical data. Synthetic EHRs offer a promising alternative: if generated records preserve key statistical and clinical properties without exposing sensitive information, they can support research, model development, system testing, and education in a privacy-preserving way.

Recent diffusion-based models such as EHDDiff (Wang et al., 2023) have demonstrated that high-fidelity and privacy-conscious EHR synthesis is achievable, showing that diffusion models can learn meaningful clinical structure directly from real-world datasets. Building on this progress, our objective is to explore whether more advanced generative techniques—specifically text-conditioned diffusion, classifier-free guidance (Ho & Salimans, 2022), and latent-space denoising (Rombach et al., 2022)—can further improve the realism, controllability, and interpretability of synthetic EHRs.

Motivated by advances in large language models and conditional generation, we propose conditioning the diffusion process on natural-language clinical descriptions generated automatically from each patient’s structured record. This yields a more flexible, controllable and semantically meaningful EHR synthesis method.

This leads to our central research question:

Can a diffusion model conditioned on clinical text descriptions generate higher-quality and more clinically faithful synthetic EHR vectors?

By investigating this, we aim to understand the potential and limitations of modern conditional diffusion methods for trustworthy, privacy-preserving healthcare data synthesis.

2 Related Work

2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion probabilistic (DDPM) (Ho et al., 2020) models provides the core theoretical framework behind model diffusion models. In this work, Ho et al break the model training process into two distinct stages. The first stage, known as the forward process, is to define a Markov process that incrementally adds Gaussian noise to a training sample over T timesteps, effectively transforming a clean vector into pure noise by the end. The second stage is the reverse process where a neural network ϵ_θ , parameterized as $\epsilon_\theta(x_t, t)$ is trained to predict the amount of noise a sample x_t contains at timestep t by minimizing the mean squared objective $\|\epsilon(x_t, t) - \epsilon\|^2$ where ϵ represents the true noise. This method achieved state of the art performance for unconditional image generation at the time of its writing and surpassed GAN-based methods to become the standard process for generating synthetic images. DDPM (Ho et al., 2020) provides the fundamental framework for data generation that we will follow by adapting their forward and reverse processes for our own task.

2.2 Elucidating Diffusion Models

EDM (Karras et al., 2022) refines diffusion modeling by introducing a continuous noise parameterization and a principled sampling procedure that improves stability and sample quality. Instead of using fixed variance schedules, EDM (Karras et al., 2022) defines noisy latents as $x_t = x_0 + \sigma\epsilon$, enabling a cleaner training objective and more consistent reverse-diffusion dynamics. These ideas have since influenced many modern diffusion systems. In our work, EDM (Karras et al., 2022) is relevant because its parameterization and sampling strategy transfer naturally to latent-space generation, providing a stable foundation for conditioning on clinical text embeddings.

2.3 Classifier-free guidance diffusion

Classifier-Free Guidance (Ho & Salimans, 2022) introduces a simple and effective way to perform conditional diffusion without training an external classifier. The model is trained jointly on conditional and unconditional inputs by randomly dropping the conditioning vector with some probability. At sampling time, conditional and unconditional predictions are combined to steer the reverse diffusion process toward the desired condition. This approach provides the foundation for our own conditional generator, where we replace class embeddings with natural-language clinical summaries.

2.4 EHRDiff

EHRDiff (Wang et al., 2023) provides a preliminary exploration into using diffusion to generate synthetic electronic health records. In this work, they adapt many of the existing diffusion techniques discussed to generate 1D vectors representing electronic health records. In our project, we attempt to extend these ideas to generate electronic health records conditioned on patient descriptions.

2.5 Stable Diffusion

Stable Diffusion (Rombach et al., 2022) introduced the now-standard idea of running the diffusion process in a learned latent space rather than the raw data domain. By encoding images into a compact representation, diffusing in that space, and decoding afterward, the model achieves far greater efficiency while preserving semantic structure.

It also popularized text-conditioned diffusion via a CLIP text encoder and token-wise cross-attention, enabling rich natural-language control. While our work similarly performs diffusion in latent space, EHR vectors lack spatial structure, so we use a dense autoencoder latent and a single global sentence embedding rather than token-level cross-attention. Still, Stable Diffusion (Rombach et al., 2022) provides the conceptual basis for latent diffusion and conditional generative control.

3 Method

3.1 Data Preprocessing

To construct the structured vectors used for diffusion training, we preprocess MIMIC-IV by converting each patient record into a unified numerical representation. From the *patients* table we extract gender, anchor age, and death indicator; from the *admissions* table we use the most recent admission to obtain marital status, race, and lifetime admission count. Categorical fields are one-hot encoded using globally consistent enumeration maps.

Diagnosis codes require normalization due to mixed ICD-9/ICD-10 usage. ICD-9 codes are mapped to ICD-10 using CMS GEMs (Centers for Medicare & Medicaid Services, 2018). All ICD-10 codes are stripped of punctuation and collapsed to their three-character category (e.g., “E1165” → “E11”), preserving broad diagnostic groups while reducing sparsity. Categories are sorted and assigned fixed indices.

Each patient’s vector is formed by setting their diagnostic categories to 1.0 and concatenating:

(1) gender, (2) age, (3) death indicator, (4) total admissions, (5) one-hot marital status, (6) one-hot race, and (7) the multi-hot diagnosis vector.

3.2 Label Generation and Embedding

To create the conditioning labels for our diffusion model, each structured MIMIC-IV vector is first converted into a deterministic text description containing only the patient’s expressed attributes: gender, age, mortality status, marital status, race, and all active diagnosis codes (or an explicit note if none are present). This ensures that the text faithfully mirrors the underlying structured vector.

For patients with diagnoses, this text block is summarized into a single clinical sentence using BioMistral-7B under a tightly constrained prompt that forbids invented conditions and enforces one-sentence output. Patients without diagnoses bypass the LLM entirely and receive a deterministic template sentence to avoid hallucinations.

The resulting summaries are encoded using *embedding-gemma-300m-medical*, producing domain-specific sentence embeddings. Each embedding is L2-normalized to ensure stable conditioning and consistent guidance strength during diffusion training.

3.3 Label Screening

After generating the raw text labels, we perform an LLM-based screening step to evaluate their correctness and quality. Each label is reviewed by Qwen2.5-14B-Instruct, which receives (1) a formatted list of the patient’s non-zero structured features and (2) the corresponding generated summary. The judge scores the label from 1–5 across five categories: clinical correctness, completeness, precision, conciseness, and formatting.

For each batch, we collect per-label scores and compute category-wise means to track labeling quality during development. High-scoring vector–label pairs can optionally be retained as a clean subset for future supervised fine-tuning of the labeling LLM.

3.4 HealthForge Diffusion Model

Our generative model follows the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) framework and incorporates classifier-free guidance (CFG) (Ho & Salimans, 2022), but adapts both components to the unique challenges of structured EHR vectors. Unlike image-based models, our diffusion process operates in the latent space of a custom autoencoder, enabling efficient denoising while preserving mixed-type feature structure.

3.4.1 Latent-Space Diffusion

Inspired by latent diffusion models (e.g., Stable Diffusion (Rombach et al., 2022)), we perform denoising directly in a 1024-dimensional latent vector produced by our EHR autoencoder. This differs fundamentally from spatial latent diffusion:

- Image LDMs (Latent Diffusion Models) use spatial latent tensors (e.g. $64 \times 64 \times 4$), enabling convolutions and region-wise consistency.
- EHR latent space is a dense, non-spatial vector, so the diffusion model relies on a series of Transformer blocks incorporating cross and self attention rather than UNets.
- Every latent dimension interacts globally, with no locality or neighborhood structure.

Let $z_0 \in \mathbb{R}^{1024}$ be the clean latent. Each step of training samples a noise level λ via $\lambda = -2 \log \tan(au + b)$ for a uniformly distributed $u \in [0, 1]$, where $b = \arctan(e^{-\lambda_{max}/2})$ and $a = \arctan(e^{-\lambda_{min}/2})$. We obtain the signal coefficient α_λ and the noise coefficient σ_λ :

$$\alpha_\lambda = \sqrt{\frac{1}{1 + e^{-\lambda}}}$$

$$\sigma_\lambda = \sqrt{1 - \alpha_\lambda^2}$$

and construct the noised latent:

$$z_\lambda = \alpha_\lambda z_0 + \sigma_\lambda \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

The model is trained to predict ϵ through a MSE learning objective, recovering the noise vector following DDPM (Ho et al., 2020) conventions.

3.4.2 Conditioning on Clinical Text Embeddings

Since the embedding is a single global vector rather than a token sequence, conditioning is performed through cross-attention between the latent vector and the text embedding, effectively globally integrating the textual signal into the latent representation. This design is specialized for non-spatial dense latents, unlike image diffusion models that require spatial cross-attention.

3.4.3 Classifier-Free Guidance

To enable controllable generation, we adopt classifier-free guidance (Ho & Salimans, 2022). During training, the text embedding is dropped with probability p :

$$\tilde{e} = \begin{cases} 0, & \text{with probability } p \\ \text{normalized embedding,} & \text{otherwise} \end{cases}$$

For each timestep t during sampling, we pass the latent vector z_t to the model ϵ_θ to predict $\epsilon_{cond_t} = \epsilon_\theta(z_t, \tilde{e})$ and $\epsilon_{uncond_t} = \epsilon_\theta(z_t, 0)$. We then compute:

$\epsilon_{\text{pred}} = (1 + s) \epsilon_{\text{cond}} - s \epsilon_{\text{uncond}}$, where s is a hyperparameter for the guidance scale.

This allows clinicians to apply natural-language control (e.g., “*elderly patient with heart failure*”) while preserving denoising stability.

3.4.4 Reverse Diffusion and Latent Reconstruction

As presented in EDM (Karras et al., 2022), the full reverse update can be decomposed into three conceptual terms:

Deterministic drift toward the previous SNR level: This part pulls the sample toward what it should look like at timestep $t - 1$:

$$\text{Drift term:} \quad \exp(\lambda_t - \lambda_{t-1}) \left(\frac{\alpha_{t-1}}{\alpha_t} z_t \right)$$

Denoised direction using the model's prediction: This term uses the denoised estimate predicted from the network’s output. We define the predicted clean latent sample x_t at timestep t as:

$$x_t = \frac{z_t - \sigma_t \epsilon_{\text{pred}}}{\alpha_t}$$

$$\text{Denoised term:} \quad (1 - \exp(\lambda_t - \lambda_{t-1})) \alpha_{t-1} x_t$$

Stochastic noise injection (variance interpolation): This term preserves the diffusion process’s randomness, where v is a fixed hyperparameter:

$$\text{Noise term:} \quad \tilde{\sigma}^{1-v} \sigma^v \xi, \quad \xi \sim \mathcal{N}(0, I)$$

Putting the three pieces together:

$$z_{t-1} = \underbrace{\exp(\lambda_t - \lambda_{t-1}) \left(\frac{\alpha_{t-1}}{\alpha_t} z_t \right)}_{\text{drift}} + \underbrace{(1 - \exp(\lambda_t - \lambda_{t-1})) \alpha_{t-1} x_t}_{\text{denoised update}} + \underbrace{\tilde{\sigma}^{1-v} \sigma^v \xi}_{\text{noise}}$$

The final latent z_0 is de-normalized using dataset-wide latent statistics and decoded through the EHR decoder, producing a structured EHR vector. Continuous dimensions (age, admissions count) are unscaled and clipped, while binary features are thresholded to maintain medical plausibility.

3.4.5 Summary of Contributions

While our approach builds on DDPM (Ho et al., 2020), CFG (Ho & Salimans, 2022), and latent diffusion concepts (Rombach et al., 2022), several technical innovations are specific to the EHR modality:

- Diffusion is performed on non-spatial clinical latent vectors, not spatial feature maps.
- Conditioning uses sentence-level medical embeddings instead of token-wise text encoders and spatial cross-attention.
- The decoder reconstructs mixed-type structured data, requiring postprocessing not present in traditional LDMs.
- The training framework aligns latent vectors with clinical summaries generated and screened through LLMs.

These adaptations enable diffusion-based generation of realistic EHR vectors conditioned on interpretable clinical descriptions.

4 Experiments

4.1 Evaluation

We evaluate our synthetic EHR generator using the core metrics from EHRDiff (Wang et al., 2023), supplemented with an additional conditional metric tailored to our model. Together, these measures assess marginal realism, multivariate structure, predictive utility, privacy preservation, and the effectiveness of text conditioning.

Because our model is conditional, several metrics differ in interpretation from prior unconditional methods (eg. (Wang et al., 2023)), so our results focus on how well the generated data reflects clinical structure and responds to text conditioning rather than on direct comparison to previous models.

4.1.1 Utility Metrics

Dimension wise prevalence correlation evaluates how well the synthetic dataset reproduces the marginal prevalence of each clinical feature in the real MIMIC-IV dataset. For every feature dimension, we compute the proportion of patients with a non-zero entry and compare the resulting prevalence vectors from the real and synthetic datasets. Their alignment is summarized by the Pearson correlation coefficient, which captures whether common diagnoses and conditions appear at similar frequencies in both datasets.

HeathForge achieves a dimension-wise prevalence correlation of 0.4347 ± 0.0007 .

This moderate positive correlation indicates that the model reproduces broad patterns of feature frequency, particularly for more common diagnoses. However, noticeable deviations remain for lower-prevalence or sparsely expressed features, suggesting room for improvement in modeling rare clinical codes.

Correlation matrix distance evaluates whether the synthetic dataset preserves the *multivariate structure* of real clinical data. Many EHR features exhibit strong co-occurrence patterns—chronic conditions cluster, certain diagnoses imply specific procedures, and demographic attributes shape risk profiles. To capture this structure, we compute the pairwise Pearson correlation matrix for real and synthetic datasets, take the elementwise absolute difference, and average across all entries to obtain a single CMD score.

HealthForge achieves a CMD of **0.1345 ± 0.0064** .

This relatively low distance indicates that the generator preserves a substantial portion of the real dataset’s feature–feature relationships. Even though marginal prevalence alignment is imperfect, the diffusion process learns many deeper co-occurrence patterns that drive clinical structure. This suggests that the model captures meaningful multivariate dependencies, which is essential for downstream predictive modeling and realistic synthetic cohort behavior.

Dimension wise prediction F1 evaluates whether models trained on synthetic data generalize to real patient data. Following EHRDiff, we select the top thirty highest-entropy features from the real dataset and treat each as a binary prediction target. For each target, we train two logistic regression classifiers—one trained on real data and one trained on synthetic data—using all remaining features as inputs. Both models are evaluated on the same real validation set, and we compute the F1 scores across all targets. The alignment between these two F1 vectors reflects how similarly synthetic and real data support downstream predictive modeling.

HealthForge obtains a dimension-wise prediction F1 alignment of **0.9195 ± 0.0010** .

This strong alignment indicates that classifiers trained on synthetic records learn nearly the same predictive structure as those trained on real data. Despite some mismatch in marginal feature prevalence, the synthetic dataset preserves the underlying relationships that matter for supervised learning. This result suggests that our conditional diffusion model captures clinically meaningful dependencies and supports realistic downstream modeling performance.

4.1.2 Privacy Metrics

Attribute inference risk evaluates whether synthetic data unintentionally leaks sensitive patient information. The attacker is assumed to know a subset of common clinical features for a patient (specifically, the 256 most frequent codes) and attempts to infer missing attributes by

retrieving nearest synthetic neighbors in feature space. For each query, the attacker performs a majority vote over these neighbors to guess the hidden attributes, and we measure the F1 score of these inferred predictions.

HealthForge yields an attribute inference F1 of 0.4972 ± 0.4967 .

The mean score remains well below 1.0, indicating that on average synthetic records do not explicitly expose hidden patient attributes. However, the high variance reveals a weakness: attack success varies substantially across patient subsets, with certain clinically structured patterns allowing noticeably better inference than others. This behavior stems from the model reproducing strong real-world correlations rather than memorizing individual records, but it still highlights a vulnerability—some attribute groups are easier to infer than intended. Overall, while the dataset offers reasonable protection, reducing variability in attribute inference risk remains an important area for improvement.

Membership inference risk tests whether synthetic samples lie close enough to real training records to let an attacker guess if a specific patient was included in the model’s training set. For each real record, we measure the distance to its nearest synthetic neighbor. If that distance falls below a learned decision threshold, the attacker predicts that the patient was present in the training data. We compute the F1 score of these predictions to quantify attack success.

HealthForge achieves a membership inference F1 of 0.1149 ± 0.1146 .

This near-random score indicates low susceptibility to membership inference attacks, suggesting that the model does not memorize individual patient records. The modest variance reflects that some real samples may incidentally lie closer to synthetic ones—often due to shared common clinical patterns—not because the generator reproduces specific training examples. Overall, the results demonstrate strong privacy protection and appropriate generalization behavior.

4.1.3 Conditional Accuracy Metric

Conditional prompt accuracy assesses whether conditioning genuinely guides generation, a metric tailored to our conditional diffusion setup. Each natural-language prompt is converted into expected feature values (demographics, multi-hot categories, and diagnosis indicators), and every generated sample is scored on how well it matches these expectations. Continuous attributes allow small tolerance, while diagnoses are evaluated as a group to accommodate clinically related conditions. A prompt’s accuracy is the mean correctness across its 100 generated samples.

HealthForge achieves a mean prompt accuracy of 0.8012 ± 0.1993 .

This indicates that the conditional generation pipeline reliably reflects prompt semantics, with most samples matching the intended demographic and diagnostic patterns. The variance suggests that prompts involving multiple or less frequent conditions are more challenging, but overall the model demonstrates strong text-to-EHR alignment.

4.2 Overall Interpretation and Results

Across the core utility metrics, HealthForge demonstrates strong realism: it captures the marginal prevalence of common features, preserves key multivariate correlations, and supports downstream predictive modeling nearly as well as real data. Privacy evaluation shows robust resistance to membership inference attacks, indicating that the model generalizes rather than memorizes individual patients. However, attribute inference poses a clear weakness—certain patient attributes become easier to infer due to overly strong real-world correlations reproduced by the model. Finally, high conditional accuracy confirms that natural-language prompts effectively guide generation. Overall, HealthForge offers realistic and controllable synthetic EHRs, with stabilizing attribute-level privacy standing out as the primary area for improvement.

5 Discussion and Further Directions

Our evaluation identifies attribute inference as the main weakness of the current system. Although membership inference remains low, certain clinical attributes can still be predicted more easily than intended. Future work should explore strategies such as privacy-regularized training losses, embedding-level noise injection, or feature-group disentanglement to reduce this leakage while preserving the overall clinical realism of the generated data.

A second direction is extending beyond static patient vectors to richer EHR representations. Incorporating laboratory measurements, medication histories, procedures, and full admission sequences would allow the model to capture temporal patterns and clinical progression, producing synthetic records that more closely resemble real patient timelines. This would also enable evaluation on more complex downstream tasks, such as event forecasting or risk-trajectory modeling.

Finally, conditional synthetic EHRs open up a range of broader applications. These include generating rare-disease cohorts for data-scarce research, producing counterfactual patient profiles to test model robustness, and creating controlled scenario datasets for benchmarking clinical algorithms under targeted conditions. Developing these applications would further demonstrate the value of text-conditioned diffusion models as practical tools for research, prototyping, and model development in settings where real EHR access is restricted.

6 Capstone Acknowledgements and Team Reflection

This work also served as our CSC490 Machine Learning Engineering capstone project. Despite being a research-oriented initiative rather than startup, the course provided valuable guidance on scoping, iteration, and constructing practical end-to-end ML systems—principles that aided in our approach nonetheless.

We adopted Infrastructure as Code (IaC) through Terraform to manage the data pipeline connecting AWS RDS and S3, reflecting the course emphasis on decoupling storage from compute. This crucially supported the numerous updates we made to data processing logic throughout development, such as to ICD normalization steps or batch preprocessing. Complementing this, we utilized managed cloud computing to facilitate novel processes that were otherwise impossible on local hardware, specifically by leveraging larger, consistent models via high-performance GPUs. In our codebase, multiple A100s and H100s ran on Modal served as the execution engine for high-intensity tasks, including data preprocessing, and label generation and screening via larger models such as BioMistral-7B and Qwen2.5-14B-Instruct.

We also applied the Transfer Learning practices emphasized in the course by integrating pretrained Hugging Face models as modular components rather than developing foundational models from scratch, such as those mentioned above as well as embedding-gemma-300m-medical. These components, used directly through the standard transformers pipeline, reduced engineering overhead and allowed us to concentrate on the novel aspects of our diffusion architecture.

The course also stressed the importance of effective literature review, a skill that played a central role in guiding our design decisions. By examining prior work—such as MedBERT (Rombach et al., 2020) and foundational guidance on synthetic data generation (e.g., Dankar & Ibrahim, 2021)—we learned to distinguish potential approaches from ideas unlikely to succeed in the structured EHR setting. This process helped identify reliable design patterns and ensured that our architecture was grounded in established research able to support our objective.

Every member of the team contributed significantly to bringing HealthForge from concept to implementation. The project required broad expertise—data engineering, distributed systems, ML research, and evaluation—and each team member contributed with the essential curiosity and initiative needed to take on these skillsets. Diffusion modeling was entirely new to our team at the start, yet through shared effort we developed a strong working understanding of both the theory and practice behind the generative models which lead this field of research.

Overall, this project reinforced a central lesson: sustained collaboration, curiosity, and interest are key drivers of innovation. The challenges encountered—technical, computational, and organizational—became opportunities for learning, and the final system reflects the integrated effort of the entire team.

References

- Centers for Medicare & Medicaid Services. (2018). *Diagnosis code set general equivalence mappings: ICD-10-CM/ICD-9-CM and ICD-9-CM/ICD-10-CM*. U.S. Department of Health & Human Services.
<https://www.cms.gov/files/document/diagnosis-code-set-general-equivalence-mappings-icd-10-cm-icd-9-cm-and-icd-9-cm-icd-10-cm.pdf>
- Dankar, F. K., & Ibrahim, M. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences*, 11(5), Article 2158.
<https://doi.org/10.3390/app11052158>
- Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. arXiv.
<https://arxiv.org/abs/2006.11239>
- Ho, J., & Salimans, T. (2022). *Classifier-Free Diffusion Guidance*. arXiv.
<https://arxiv.org/abs/2207.12598>
- Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). *Elucidating the Design Space of Diffusion-Based Generative Models*. arXiv. <https://arxiv.org/abs/2206.00364>
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2020). *Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction*. arXiv. <https://arxiv.org/abs/2005.12833>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. arXiv. <https://arxiv.org/abs/2112.10752>
- Wang, H., Hotchkiss, J., & Zhang, J. (2023). *EHRDiff: Exploring Realistic EHR Synthesis with Diffusion Models*. arXiv. <https://arxiv.org/abs/2303.05656>