Video-based Disease Progression Simulation

Xu Cao XUCAO2@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Kaizhao Liang KAIZHAOL@UTEXAS.EDU
University of Texas at Austin, USA

Kuei-Da Liao CRAIGKDLIAO@GMAIL.COM
Meta, USA

Tianren Gao TERRYGAO87@BERKELEY.EDU
Meta, USA

Zhiguang Ding 15622134805@QQ.COM
Shenzhen Nanshan People's Hospital, China

Jianguo Cao JIANGUOCAO@PEDIAMED.AI
Shenzhen Children's Hospital, China

Zheng Chen CHENZ@SANKEN.OSAKA-U.AC.JP
Osaka University, Japan

Jintai Chen JINTAICHEN@HKUST-GZ.EDU.CN
Hong Kong University of Science and Technology (Guangzhou), China

James M. Rehg JREHG@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Jimeng Sun JIMENG@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Abstract

Modeling disease progression is crucial for improving the quality and efficacy of clinical diagnosis and prognosis, but it is often hindered by a lack of longitudinal medical image monitoring for individual patients. To address this challenge, we propose MedDream, the first video-based disease progression framework that enables controlled manipulation of disease-related image and video features, allowing precise, and personalized simulations of disease progression. Our approach begins by disease trajectory description recaptioning. Next, a controllable multi-round diffusion model simulates the disease progression state for each patient, creating realistic intermediate disease state sequences. Finally, a diffusion-based video transition generation model interpolates disease progression between these states. We validate our framework across three medical imaging domains: chest X-ray, fundus photography, and skin image. Our results demonstrate that MedDream significantly outperforms baseline models in generating coherent and clinically plausible disease trajectories. Two user studies by veteran physicians, provide further validation into the clinical relevance of the generated sequences. MedDream has the potential to assist healthcare providers in modeling disease trajectories, interpolating missing medical image data, and enhancing medical education through realistic, dynamic visualizations of disease progression.

Data and Code Availability This paper uses publicly available medical image data from the CheXpert Plus, MIMIC-CXR-JPG, ISIC 2024, ISIC 2018, Kaggle Diabetic Retinopathy Detection Challenge. The code used to preprocess the data, train the models, and run the experiments is available at github.com/PediaMedAI/PIE.

Institutional Review Board (IRB) Since all the datasets used in this study are publicly available, this study does not require IRB approval for dataset usage. For user study in the evaluation pipeline, we got IRB approval from the partner hospital.

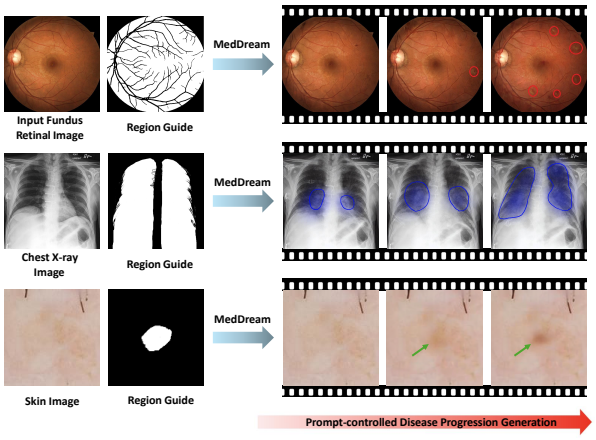

Figure 1: Illustrative examples of video-based disease progression simulation (6-8s) using predefined medical reports and our proposed method. The top sequence depicts a patient's Diabetic Retinopathy. The middle sequence demonstrates the Edema in a patient's lung. The bottom sequence demonstrates the Benign Skin Lesion in a patient's skin.

1. Introduction

Disease progression refers to the way an illness evolves in an individual over time. Understanding this progression enables healthcare professionals to develop effective treatment strategies, anticipate complications, and adjust care plans accordingly. Disease progression modeling can also be seen as a form of human digital twin, laying the foundation for future precision medicine (Li et al., 2024c; Tang et al., 2024; Vallée, 2024). However, modeling disease progression on medical images presents significant challenges. These challenges arise primarily from the lack of continuous monitoring of individual patients over time, as well as the high cost and risks associated with collecting longitudinal imaging data (Liu et al., 2015; Cook and Bies, 2016; Severson et al., 2020). The intricate and multifaceted dynamics of disease progression, combined with the lack of comprehensive and continuous image or video data of individual patients, result in the absence of established methodologies for medical imaging trajectories simulation (Lee et al., 2019).

Recent advancements in image/video/3D generation models and world models present promising opportunities for simulating realistic medical videos, potentially enriching existing databases and addressing data limitations for medical world models (Li et al., 2024b; Liu et al., 2025; Temsah et al., 2025). To incorporate generative models into disease progression simulations, we establish **three** key criteria that medical video generation models must meet: (1) Models should generate videos presenting long disease progression under zero-shot setting, as there are no existing datasets for video-based disease progression at current stage. (2) The generated disease states must be semantically relevant to the initial input image. (3) The generated progression should be clinically verified and consistent with the corresponding text descriptions.

In this work, we propose MedDream, a video generation framework for simulating disease progression that integrates text inference, progressive image generation, and video clip transition generation. Specifically, our approach uses GPT-4 (Achiam et al., 2023) to recaption clinical reports and generate prompts, which are then used to progressively control disease-related features extracted by a text encoder. This approach allows us to conditionally simulate disease progression in the visual domain without significantly altering the core features of the initial image (see Figure 1). After generating a sequence of disease state images, we utilize a video transition generation model, guided by conditional masks, to interpolate between successive disease states, thereby creating a realistic simulation of disease progression. In summary, our framework is built on the invertibility of denoising diffusion probabilistic models (Ho et al., 2020; Song et al., 2020a), the visual-language alignment capabilities of context encoders (Esser et al., 2024), and frame-level synthesis.

Our theoretical analysis further demonstrates that the multi-step disease state simulation module of MedDream can be understood as a gradient descent process toward maximizing the log-likelihood of the given text conditioning. The learning rate in this iterative process decays exponentially with each forward step, allowing the algorithm to effectively explore the solution space while balancing convergence speed and stability. This guarantees that our framework moves toward the target disease manifold, ensuring that the modifications made are clinically plausible and remain bounded for medical concepts.

The contributions are summarized as follows:

- We propose the first medical video simulator MedDream for disease progression, which allows for a precise understanding of disease-related image features and leads to accurate and individualized longitudinal disease progression simulation.
- We provide theoretical evidence that our iterative refinement process is equivalent to gradient

descent with an exponentially decaying learning rate, which helps to establish a deeper understanding of applying diffusion-based generative models in healthcare research.

- We demonstrate the superior performance of MedDream over baselines in disease progression prediction with three medical domains via different evaluation metrics and physician user preference study.
- In the follow-up user study, 35 physicians agree that 76.2% of disease state sequences simulated by MedDream closely matched physicians' expectations, indicating our generation results are high related to the clinical context.

2. Related Works

Disease Progression Simulation. Longitudinal disease progression data derived from individual electronic health records offer an exciting avenue to investigate the nuanced differences in the progression of diseases over time (Schulam and Arora, 2016; Stankeviciute et al., 2021; Mikhael et al., 2023). However, most of the previous works are based on HMM (Wang et al., 2014; Liu et al., 2015) or deep probabilistic models (Alaa and van der Schaar, 2019) without using data from imaging space. Some recent works have started to resolve image disease progression simulation by using deep-generation models. (Ravi et al., 2022; Jung et al., 2023) utilized the Generative Adversarial Networks (GANs) based model and linear regressor with individual sequential monitoring data for Alzheimer's disease progression simulation in MRI imaging space. All of these methods have to use full sequential images as training sets and are hard to adapt to the general medical imaging domain due to lack of sequential data.

Generative Models. Denoising Diffusion Models (Ho et al., 2020; Song et al., 2020a; Rombach et al., 2022; Karras et al., 2022) have become increasingly popular due to their ability to create high-resolution realistic images from textual descriptions. Among the various text-to-image models, latent diffusion model (LDM) (Rombach et al., 2022; Esser et al., 2024) and its follow-up image-to-image editing works (Brooks et al., 2022; Parmar et al., 2023; Orgad et al., 2023) has received considerable attention because of its impressive performance in generating high-quality images and its ability to edit scenarios across multiple modalities. While image generation has seen substantial progress in general domains, its application in the medical field remains less explored (Yi et al., 2019). Earlier work using Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and GANs (Goodfellow et al., 2020) focused on generating medical images like X-rays and MRIs to address the issue of limited training data (Costa et al., 2017; Zhang et al., 2018; Madani et al., 2018; Nie et al., 2017). The introduction of LDMs significantly improved the quality of these images (Packhäuser et al., 2023; Kazerouni et al., 2023; Müller-Franzes et al., 2023), even extending to 3D synthesis (Khader et al., 2023; Dar et al., 2023). Recently, efforts have been made to unify medical report generation with image synthesis (Lee et al., 2023; Bluethgen et al., 2024), and design image editing pipeline for counterfactual medical image generation (Gu et al., 2023; Kyung et al., 2024).

Video Generation Models in Medial Domain. Text-to-video models have attracted significant attention from both academia and industry (Sun et al., 2024a), leading to the development of video generation models and world simulators such as Sora (Midjourney, 2024; Zhu et al., 2024b), Pika (Pika, 2024), and Stable Diffusion Video (Blattmann et al., 2023a). The core of these models often involves fine-tuning or integrating additional modules or priors into pre-trained text-to-image diffusion models using video data, as seen in Make-A-Video (Singer et al., 2022), PYoCo (Ge et al., 2023), and LaVie (Wang et al., 2023), SEINE (Chen et al., 2023), AnyV2V (Ku et al., 2024). The integration of these models into healthcare holds immense potential (Sun et al., 2024b; Khader et al., 2023), enhancing diagnostic and surgical decision-making while improving generalizability. Recent medical video generation efforts include producing surgical videos (Cho et al., 2024; Chen et al., 2024), and creating text-to-video simulations for diverse imaging modalities (Li et al., 2024a; Wang et al., 2024; Kurt et al., 2025; Wang et al., 2025). However, these methods present many challenges, particularly because video data for the medical domain is hard to collect (Kang et al., 2024). To fill this gap, we explore the potential of applying text-to-image generation models to simulate disease progression videos in zero-shot manner.

3. Problem Statement

Video generation models need to be trained with a large amount of text-to-video or image-to-video data.

However, it is almost impossible to obtain large-scale longitude medical imaging data (can be also considered as a type of medical video data) as most patients may not go to the same hospital for follow-up treatment and the hospitals also lack medical imaging and clinical reports in the early stages of diseases.

In our paper, we reconsider this problem in another way. Assume we have a text-to-image generation model pretrained a large amount of clinical report to medical image data. Given an input medical image x_0, and clinical report and medical history label pair (y_0 and y_N) at time step 0 and time step N, where $N+1$ is the total number of states of the predicted disease. The predicted disease progression is a video sequence Z, which can be separated by a set of short video clips $\{z_0, z_1, z_2, ..., z_{N-1}\}$, where $z_i \in \mathbb{R}^{K\times H\times W\times C}$ is a video clip between disease image state x_i and x_{i+1}. K, H, W, C denote the number of frames, height, width, and channels of the video clip. K is a very small number to control the disease progression change in a limited medical imaging space. In z_i, the starting frame $x_i \in \mathbb{R}^{H\times W\times C}$ is the initial disease state and end frame $x_{i+1} \in \mathbb{R}^{H\times W\times C}$ is the end disease state.

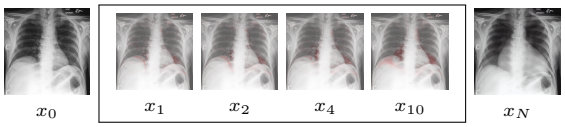

Figure 2: Visualization for cardiomegarly disease state absolute difference heatmap. The highlighted portion from x_1 to x_{10} illustrates the progression of the pathology at each step.

We separate the disease progression video generation into a two stage strategy. In the first stage, the key idea is to generate discrete disease progressive states $\{x_0, x_1, x_2, ..., x_N\}$ with medical report y:

$$x_n = f_\theta(x_{n-1}, y) \tag{1}$$

y is the text feature extracted by y_0 and y_N. In the training phase of the first stage, f_θ is a denoising diffusion model learned from independent identically distributed $(\boldsymbol{x}, y)$ from different patients.

In the second stage, we adopt video latent diffusion models finetuned with video data in the general domain. In doing so, we convert the disease progression video generation task into a frame-level transition generation problem:

$$z_i = g_\phi(x_i, x_{i+1}) \tag{2}$$

The output videos $\{z_0, z_1, z_2, ..., z_{N-1}\}$ finally concatenate into the disease progression video $Z \in \mathbb{R}^{(N(K-1)+1)\times H\times W\times C}$.

4. Method

As shown in Figure 3, MedDream contains two main components: (i) **Progressive disease image editing (PIE)** with medical domain-specific diffusion model (Top of Figure 3), and (ii) **Transition Generation Process** between generated disease states with video latent diffusion model (Bottom Left of Figure 3). The output sequences of (ii) are concatenated to the final video output (Bottom Right of Figure 3).

The first component PIE is a long-sequence medical image editing framework proposed to refine and enhance images iteratively and discretely, allowing clinical report-based prompts for precise adjustments to simulate disease development while keeping realism. Unlike traditional image editing techniques, PIE involves a multi-stage process where each step builds upon the previous one, intending to achieve a final result that is more refined than if all changes were made at once. Transition generation is used in the long video generation model to connects different narrative moments. Once the frame-level sequence is generated by PIE, we will provide each pair of adjacent frames and use transition prompts and disease region mask to control the style and content, creating intermediate frames that further illustrate the transition or progression within the medical video sequence.

4.1. Progressive Image Editing (PIE)

Procedure. The inputs to PIE are a discrete medical image x_0^0 depicting any start or middle stage of a disease and a corresponding terminal stage prompt y_N inferred by medical doctor for clinical education purpose and then re-captioned to standard clinical report format by GPT-4o (Achiam et al., 2023), providing the potential hint of the patient's disease progression. During training, all text data follows the same format of y_N. The Latent y will be the text conditioning of the diffusion model (Rombach et al., 2022). y is generated from a pretrained text encoder from CLIP (Radford et al., 2021) (clip-vit-large-patch14), where the text input is y_N. The output generated

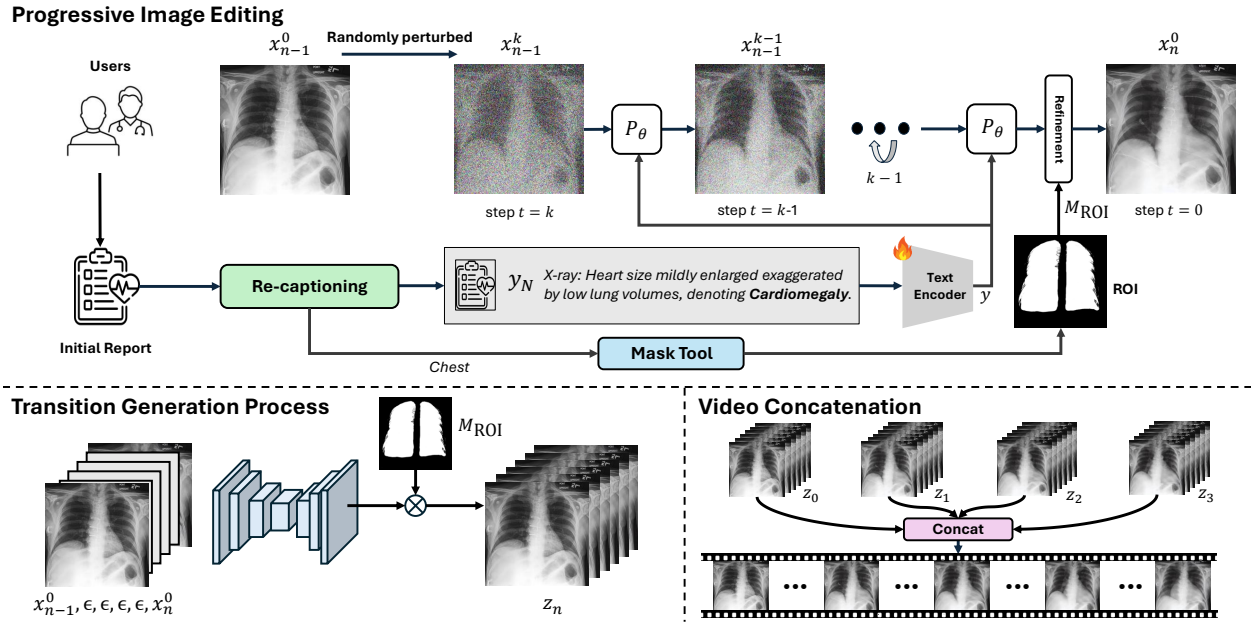

Figure 3: Overview of the MedDream inference pipeline. The above part denotes the single step of PIE. For any given step n in PIE, we first utilize inversion of diffusion model to procure an inverted noise map. Subsequently, we denoise it using GPT-4o re-captioned clinical reports from the future state and use the ROI mask to refine the editing after the last step of denoising. The output of a single step of PIE is the input for the next step $n+1$, thus ensuring a gradual and controllable disease progression simulation. After simulating N steps, the image is converged to the final state. The below part shows the transition generation process between disease states. We use ROI mask to control the mask recovery of SEINE and finally output the long sequence of video-based disease progression.

by PIE is a sequence of images presenting the disease progression, $\{x_0^0, x_1^0, \cdots, x_N^0\}$. The iterative PIE procedure is defined as follows:

Proposition 1 * *Let $x_n^0 \sim \chi$, where χ is distribution of photo-realistic medical images, y be the text conditioning, running $\text{PIE}_n(\cdot,\cdot)$ recursively is denoted as following, where $n = \{N, N-1, \cdots 1\}$,*

$$x_n^0 = \text{PIE}_n(x_{n-1}^0, y) \tag{3}$$

$$x_N^0 = \underbrace{\text{PIE}_N \circ \text{PIE}_{N-1} \circ \cdots \circ \text{PIE}_1}_{N \text{ times}}(x_0^0, y) \tag{4}$$

Then, the resulting final output x_N^0 maximizes the posterior probability $p(x_N^0 |\, x_0^0, y)$.

To run the inference pipeline of PIE to generate a discrete disease progression image sequence, we use the original input image x_0^0 as the start point. The hyperparameters are the number of progression stage N, diffusion steps T, text conditional vector y, noise strength γ determines the number of noise steps to add, diffusion parameterized denoiser ϵ_θ, and a region of interest (ROI) mask M_{ROI}, where each pixel in $M_{\text{ROI}}^{i,j} \in [0,1]$.

Since PIE is a recursive image-to-image editing process, at progression stage n, x_{n-1}^0 is the input image. k gaussian noise is added to x_{n-1}^0 (randomly perturbed step in Figure 3).

From diffusion step k to 1,

$$x' \leftarrow \sqrt{\alpha_{t-1}}\left(\frac{x' - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x', y)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta^{(t)}(x', y) \tag{5}$$

where x' in step k is x_{n-1}^0, k is $\gamma \cdot T$, $\epsilon_\theta^{(t)}(x', y)$ is the noise prediction by denoisor, where θ is the parameter in the denoiser, t is the step from k to 0. After the last step, we use the M_{ROI} initially generated by

*The proof of Proposition 1 and Proposition 2 are shown in the supplementary material.

pretrained medical segmentation foundation models and then slightly edit by human to control and refine the final output:

$$x' \leftarrow (\beta_1 \cdot (x' - x_0^0) + x_0^0) \cdot (1 - M_{\text{ROI}}) + (\beta_2 \cdot (x' - x_0^0) + x_0^0) \cdot M_{\text{ROI}} \quad (6)$$

where β_1, β_2 are hyperparameter to control the interpolation between generated result and the input image. The last output image x' is x_{n-1}^T, which is also the input x_n^0 of the next step ($n+1$ step) disease state generation. Equation 6 guarantees the editing is regional based and avoids the image distortion caused by multiple times image editing. It is worth noting that Equation 6 can generalize to arbitrary diffusion backbones (Rombach et al., 2022; Esser et al., 2024).

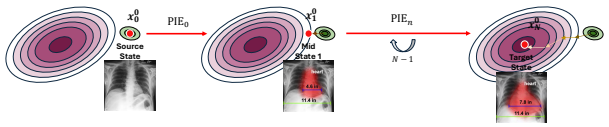

Figure 4: Simulating a sequence of disease stages using MedDream. The red dots indicate the current stage, while the green and purple contours represent the distributions of the image-based health stage and disease stage, respectively. Beginning with an image at the health stage, we perturb it with Gaussian noise and progressively remove the noise, arriving at an intermediate stage (single step of PIE). We then run PIE recursively until the image no longer changes. This iterative process gradually projects the features of the future disease state onto the manifold of the original images.

With each round of editing as shown in the Figure 4, the image gets closer to the objective by moving in the direction of $-\nabla \log p(x|y)$. The step size would gradually decrease with a constant factor. The iterative convergence analysis is as follows:

Proposition 2 *Assuming* $\|x_0^0\| \leq C_1$ *and* $\|\epsilon_\theta(x, y)\| \leq C_2$, $(x, y) \in (\chi, \Gamma)$, *for any* $\delta > 0$, *if*

$$n > \frac{2}{\log(\alpha_0)} \cdot (log(\delta) - C) \quad (7)$$

then,

$$\|x_{n+1}^0 - x_n^0\| < \delta \quad (8)$$

where, $\lambda = \frac{\sqrt{\alpha_0 - \alpha_0\alpha_1} - \sqrt{\alpha_1 - \alpha_0\alpha_1}}{\sqrt{\alpha_1}}$, χ *is the image distribution,* Γ *is the text condition distribution ,* C_1 *and* C_2 *are two constants.* $C = \log((\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2)$

Proposition 2 shows as n grows bigger, the changes between steps would grow smaller. Eventually, the difference between steps will get arbitrarily small. It guarantees modifications to any medical imaging inputs are bounded by a constant, which avoids generating unreasonable output images. The proof of Proposition 2 is shown in the supplementary material.

4.2. Transition Generation Process

The concept of scene transition generation is first proposed by SEINE (Chen et al., 2023), which is a short-to-long video diffusion model. In MedDream, we use M_{ROI} to control SEINE to connect the disease progression between each step generated by PIE,

$$z'_n = \text{Concat}(x_{n-1}^0, \underbrace{\epsilon, \cdots, \epsilon}_{\text{random noise}}, x_n^0) \quad (9)$$

$$z_n = \frac{x_{n-1}^0 + x_n^0}{2} \cdot (1 - M_{\text{ROI}}) + g(z'_n) \cdot M_{\text{ROI}} \quad (10)$$

,where z_n is a video clip with the first and last frames are the input x_{n-1}^0 and output x_n^0 from progression stage n in PIE. Between x_{n-1}^0 and x_n^0, all frames are masks with random noise. By predicting and modeling the noise, the transition generation process $g(\cdot)$ aims to extend realistic, visually coherent transition frames that seamlessly integrate the visible frames with the unmasked ones. After generate all video clips between disease states, we concatenate them to generate the final long disease progression video ouput:

$$Z = \text{Concat}(z_0, \cdots, z_{N-1}) \quad (11)$$

4.3. Model Training

We finetuned domain-specific text-to-image Stable Diffusion model on the training set of three different types of disease. The datasets used in the model training are CheXpert Plus (Chambon et al., 2024) and MIMIC-CXR (Johnson et al., 2019) for chest X-ray classification and report generation (Irvin et al., 2019; Chambon et al., 2024; Johnson et al., 2019), ISIC 2024 and ISIC 2018 (Codella et al., 2019; Tschandl et al., 2018; Kurtansky et al., 2024), and Kaggle Diabetic Retinopathy Detection Challenge (CHF, 2015). Each of these datasets presents unique challenges and all of them having large-scale of data, making them suitable for testing the robustness and versatility of our

proposed method. We also collected over 50 healthy data among the test set from these datasets as initial input data for disease progression video generation. These data were used for disease progression simulation. Three groups of progression visualization results can be found in Figure 6.

To fine-tune the Stable Diffusion model (stable-diffusion-v1.4), we center-crop and resize the input X-ray images to 512×512 resolution, fundus retinal images to 512×512 resolution, skin images to 128×128 resolution. For each image, we also change the context prompt with a start sentence "[X-ray OR Fundus Retinal Image OR Skin Image]" and then connect it with the disease class and original clinical report. The loss function to finetune the denoiser is the L2 loss. Then we utilize the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay set at 0.01, and set the first 500 training steps as warm-up steps. Additionally, we employ a cosine learning rate scheduler (Loshchilov and Hutter, 2016), with the base learning rate set at 5×10^{-5}. All models undergo fine-tuning for 20,000 steps on 4 NVIDIA H100 GPUs, with each GPU handling a batch size of 8.

5. Results

In this section, we present experiments on various disease progression tasks. Experiments results demonstrate that MedDream can simulate the disease-changing trajectory that is influenced by different medical conditions. Notably, MedDream also preserves unrelated visual features from the original medical imaging report, even as it progressively edits the disease representation. Figure 6 showcases a set of disease progression simulation examples across three distinct types of medical imaging.

Methods	GPT4o$_{score}$ (↑)	Conf$_{final}$ (↑)	Conf$_{seq}$ (↑)
Extrapolation (Han et al., 2022)	1.00	0.054	0.032
SVD (Raw, 2023)	0.12	0.389	0.255
Our	0.96	0.712	0.681

Table 1: Simulation Experiment for chest X-ray. Each input image is from the healthy state and will be used to generate the progression of Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion. All results in this table are average score among five diseases.

Methods	GPT4o$_{score}$ (↑)	Conf$_{final}$ (↑)	Conf$_{seq}$ (↑)
Extrapolation (Han et al., 2022)	0.94	0.074	0.068
SVD (Raw, 2023)	0.00	0.121	0.092
Our	0.92	0.807	0.702

Table 2: Simulation Experiment for retinal fundus image. Each input image is from the healthy state and will be used to generate the progression of diabetic retinopathy.

Methods	GPT4o$_{score}$ (↑)	Conf$_{final}$ (↑)	Conf$_{seq}$ (↑)
Extrapolation (Han et al., 2022)	0.62	0.226	0.198
SVD (Raw, 2023)	0.00	0.201	0.104
Our	0.96	0.694	0.496

Table 3: Simulation Experiment for skin image. Each input image is from the healthy state and will be used to generate the progression of Melanocytic nevus.

5.1. Experimental Setups

Implementation Details. Stable Diffusion checkpoints (CompVis/stable-diffusion-v1-4) is used as the initual weight to finetune the Stable Diffusion model with three three medical domain. The weight for transition generation model is from SEINE (Chen et al., 2023). The mask tool used in our task is supervised finetuned with 10 image-mask pair for each medical domain. Our code and checkpoints will be publicly available upon publication. All experiments are conducted on 4 NVIDIA H100 GPUs (for training) and 1 RTX 4090 GPU (for inference).

Baselines. To our knowledge, there are no existing generation models specifically designed for simulating discrete disease progression sequences or videos under the no trainable sequential data setting. To underscore the unique strengths of MedDream, we compare it against with related baseline multi-stage diffusion generation strategy. One of them is Stable Video Diffusion (SVD), also called Stable Diffusion Walk (Raw, 2023) for short video generation. SVD is the basic of the latent-based video generation methods like Stable Diffusion Video (Blattmann et al., 2023b; Wu et al., 2022), but it do not need any training from video datasets. Another one is the Style-Based Manifold Extrapolation (Extrapolation) (Han et al., 2022) for generating progressive medical imaging with GAN, as it don't need diagnosis labeled data (Ravi et al., 2019; Han et al., 2022), which is similar to our definition setting but it need plenty of progression inference prior. In Figure 5, we showcase how these model edit the image with multi-step by prompt guidance

Method A	Method B	X-ray	Skin	Retinal
		Win Rate (↑)	Win Rate (↑)	Win Rate (↑)
Pika (Pika, 2024)	PixVerse (PixVerse, 2024)	0.42	0.50	0.54
	CogVideoX (Yang et al., 2024)	0.46	0.42	0.67
	MedDream (Our)	0.20	0.33	0.33
PixVerse (PixVerse, 2024)	Pika (Pika, 2024)	0.58	0.50	0.46
	CogVideoX (Yang et al., 2024)	0.58	0.58	0.58
	Our	0.23	0.37	0.37
CogVideoX (Yang et al., 2024)	Pika (Pika, 2024)	0.54	0.58	0.33
	PixVerse (PixVerse, 2024)	0.42	0.42	0.42
	Our	0.17	0.33	0.20
Our	Pika (Pika, 2024)	**0.80**	**0.67**	**0.63**
	PixVerse (PixVerse, 2024)	**0.77**	**0.63**	**0.67**
	CogVideoX (Yang et al., 2024)	**0.83**	**0.67**	**0.80**

Table 4: User preference A/B test from 30 verified clinicians, radiologists of the generated disease progression videos from MedDream and three SOTA image-to-video generation models.

in the manifold. During the comparison, all trainable baseline methods are using the same Stable Diffusion finetuned weights in specific dataset and also applied the same M_{ROI} for region guidance.

5.2. Evaluation

The evaluation of generated disease progression images focuses on two key aspects: alignment with the intended disease features and preservation of patient identity. To assess these aspects, we employ three primary metrics in the experiments: the GPT4o$_{score}$, Conf$_{final}$, and Conf$_{seq}$. To select the best hyperparameter, we also use traditional image generation metrics in the ablation studies.

- **GPT-4o Identity Score** (GPT-4o$_{score}$) serves as an automated sanity check to evaluate the preservation of patient identity between the initial image x_0 and the generated terminal stage x_N. Rather than acting as a rigorous clinical biometric metric, this score is specifically designed to filter out samples with severe structural distortions introduced during the progressive editing process. To ensure the reliability of this filtering task, we employ few-shot in-context learning within the prompt, guiding GPT-4o to assess whether the second image maintains the anatomical and identity-defining characteristics of the first. The output is a binary score GPT-4o$_{score} \in \{0, 1\}$ provided in a structured JSON format.

- **Final state confidence score** is based on a disease classifier f_θ. The classifier is a supervised deep network trained for binary classification between negative (healthy) and positive (disease) samples. It is defined as Conf$_{final}$ = $Sigmoid(f_\theta(x))$ and measures how well the generated images align with the target disease state. We utilize the DeepAUC maximization method (Yuan et al., 2021) with as the classifier, which is recognized for its SOTA performance on CheXpert and ISIC 2018 task 3.

- **Sequence confidence score** is based on Spearman's rank correlation coefficient to measures how well the generated image sequence align with the disease progression change within time. It is defined as Conf$_{seq}$ = $(1 - \frac{6\sum_{k=0}^{N}(r(f_\theta(x_k)) - r(k))^2}{N(N^2-1)}) * Sigmoid(f_\theta(x))$, where f_θ is the disease classifier based on DeepAUC; $r(.)$ is the rank in the sequence; N is the total steps of PIE. Conf$_{seq}$ reflect whether the disease state of the generated sequence follows the right direction and path.

- **Kernel Inception Distance (KID)** is like the Fréchet Inception Distance (FID) and also uses the Inception network. It computes the squared Maximum Mean Discrepancy (MMD) between real and generated images. Lower KID values mean greater similarity, with zero as the best (lowest) score. Unlike FID, KID is unbiased and thus more reliable when there are fewer test images than the dimensionality of the Inception

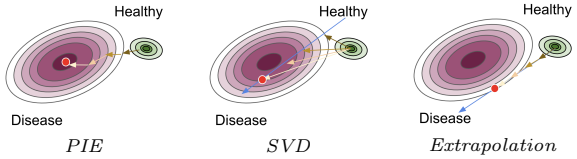

Figure 5: Understanding the progression generation path of PIE, SVD, and Extrapolation in the manifold.

features. However, both FID, KID may not reflect perceived disease image quality, making it less suitable for our tasks. We only attach KID in the ablation study.

- **Win Rate** is a clinician user study metric derived from an A/B test. In this test, clinicians invited by our clinical co-authors evaluate 36 comparative video pairs by answering a questionnaire. The Win Rate reflects user preferences regarding the realism and disease relevance of the generated videos.

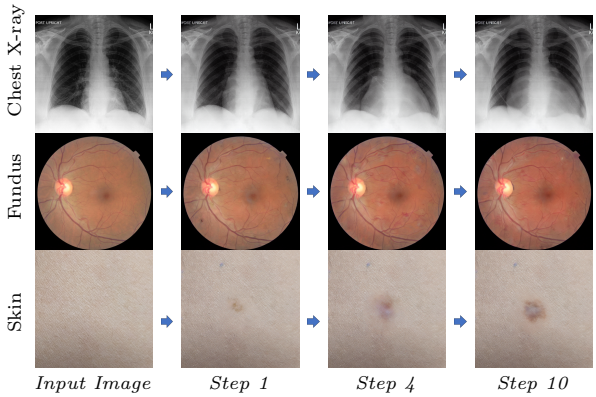

Figure 6: Disease Progression Simulation of MedDream. The top progression is for Cardiomegarly. The middle progression is for Diabetic Retinopathy. The bottom progression is for Melanocytic Nevus.

5.3. Disease State Simulation

To demonstrate the superior performance of MedDream in disease progression simulation compared to single-step editing methods, we conduct experiments on the three previously mentioned datasets. For each disease, we select 50 healthy test samples as the starting point and run MedDream, SVD, and Extrapolation with five random seeds. This process generates at least 50 disease progression imaging trajectories for each patient.

Tables 1, 2, and 3 present the main experimental results across three distinct medical imaging domains. MedDream consistently outperforms both SVD and Extrapolation in terms of final-state confidence scores and sequence-based confidence scores, while maintaining high GPT-4o identity scores. Figure 7 further illustrates qualitative results of disease progression simulation for Edema in chest X-rays using CheXpert clinical report prompts. MedDream produces more realistic and clinically coherent progressive edits compared to the baselines. While Extrapolation preserves identity well, it sacrifices effective editing, resulting in low confidence scores. In contrast, SVD introduces significant alterations in the initial step but struggles to maintain a proper progression trajectory, often generating uncontrollable noise after a few steps. A manifold-based explanation of these behaviors is provided in Figure5. Additional qualitative results across different datasets are included in the Supplementary Material.

5.4. Disease Progression Video Simulation

Table 4 shows the A/B test results between MedDream and three image-to-video generation baselines. We did not compare our method with text-to-video generation models like Stable Diffusion Video (Blattmann et al., 2023a), as these models do not support video generation from an initial medical image. Compared to PixVerse (PixVerse, 2024), CogVideoX (Yang et al., 2024), and Pika (Pika, 2024), our method demonstrates significantly higher clinician preference, achieving average win rates of 79%, 70%, and 66% for Cardiomegaly in chest X-ray, diabetic retinopathy, and benign skin lesion disease progression simulations, respectively.

5.5. Ablation Study

Module	Conf_{final}		
	Chest X-Ray	Fundus Retinal Image	Skin Image
MedDream w/o mask	0.729	0.163	0.666
MedDream with mask	0.712	0.807	0.453

Table 5: Ablation study for mask, w/o mask guidance comparisons.

Effect of Region Guide Masks. To assess the impact of segmentation masks, we conducted two ablation studies. As shown in Table 5, we compare the final-state image classification confidence scores with

Mask Module	Conf_{final}	CLIP-I	KID (↓)
Med-SAM (Ma et al., 2024)	0.978	0.962	0.142
Med-SAM-2 (Zhu et al., 2024a)	0.974	0.964	0.133
BioMedParse (Zhao et al., 2024)	0.979	0.960	0.143

Table 6: Ablation study on segmentation foundation models for M_{ROI}.

Strength	Conf_{final}	CLIP-I	KID (↓)
0.1	0.120	0.969	0.0638
0.2	0.273	0.969	0.0885
0.4	0.746	0.965	0.1142
0.5	0.978	0.962	0.142
0.6	0.995	0.956	0.1549
0.8	0.999	0.951	0.1629

Table 7: Ablation study on Strength γ selection for $N = 10$. Results indicate that strength control the stride for progressive editing.

Step (N)	Conf_{final}	CLIP-I	KID (↓)
1	0.491	0.965	0.094
5	0.881	0.963	0.121
10	0.978	0.962	0.142
50	0.975	0.962	0.130
100	0.959	0.962	0.115

Table 8: Ablation study on simulation steps N selection with $\gamma = 0.5$. Chest X-ray disease progression converge with step=10.

and without mask guidance. While removing mask guidance slightly increases Conf_{final} for Chest X-ray disease simulation, it fails for fundus retinal and skin images. Additionally, Table 6 presents experiments evaluating different medical mask segmentation tools. The results indicate that these medical segmentation foundation models do not significantly differ in their effectiveness for our task.

Effect of PIE hyperparameter. To evaluate the impact of PIE hyperparameters, we conducted three ablation studies in Table 7, 8, and 9. The results indicate that PIE will converge at progression stage $N = 10$ and the Conf_{final} of the disease state remains stable after step 10. Increasing β_2 affects the output image quality, while adjusting β_1 and β_2 does not significantly impact the CLIP-I score. In our main experiment, we choose the hyperparameter $N = 10$, $\gamma = 0.6$, $\beta_1 = 0.01$, $\beta_2 = 0.75$.

β_1	β_2	Conf_{final}	CLIP-I	KID (↓)
0.01	1.0	0.954	0.946	0.133
0.01	0.75	0.977	0.948	0.140
0.01	0.5	0.554	0.965	0.090
0.1	1.0	0.960	0.965	0.126
0.1	0.75	0.976	0.962	0.140
0.1	0.5	0.554	0.962	0.089
0.2	1.0	0.963	0.947	0.134
0.2	0.75	0.977	0.964	0.137
0.2	0.5	0.556	0.962	0.089

Table 9: Ablation study on the interpolation hyperparameters β_1 and β_2 selection shows that increasing β_2 affects the output image quality, while adjusting β_1 and β_2 does not significantly impact the CLIP-I score.

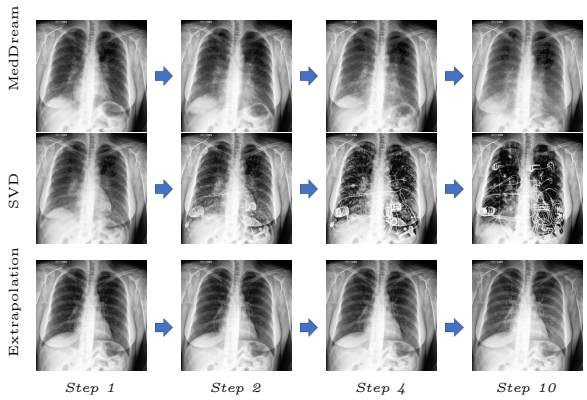

Figure 7: Using MedDream, SVD, Extrapolation to simulate Edema progression with clinical reports as input prompt.

5.6. User Study

To further assess the quality of our generated disease state sequences, we conducted another comprehensive user study from 35 physicians and radiologists with 14.4 years of experience on average to answer a questionnaire on chest X-rays. The questionnaire includes disease classifications on the generated and real X-ray images and evaluations of the realism of generated disease progression video of Cardiomegaly, Edema, and Pleural Effusion. More details of the questionnaire and the calculation of the statistics are presented in Supplementary Material. The participating physicians have agreed with a confidence of **76.2**% that MedDream simulated disease state progressions on

the targeted diseases fit their expectations. One plausible explanation is due to the nature of MedDream, the result of running progressive image editing makes pathological features more evident. The aggregated results from the user study demonstrate our framework's ability to simulate disease progression to meet real-world standards.

6. Discussion

The concept of healthcare world simulators is highly anticipated, yet healthcare data is far more constrained than in other domains (Algethami et al., 2025; Liu et al., 2025; Waisberg et al., 2024). MedDream represents one new approach in this space. However, it is crucial to emphasize that the disease progression videos generated by MedDream is educational only and may not reflect actual patient trajectories. Our ultimate goal is to develop a video generation method to augment disease progression data and interpolate missing medical images in longitudinal electronic health records (EHRs), analogous to counterfactual medical image generation. The patient is final state is known when testing MedDream. With clinician or radiologist verification, MedDream's outputs can augment existing datasets and support education of junior clinicians, facilitate communication between patients and medical doctors. As a data augmentation and simulation method, MedDream should not be used to predict future states for real clinical practice.

7. Conclusion and Outlook

In conclusion, our proposed framework, MedDream for disease progression simulation, holds great potential as a tool for medical research and clinical practice in simulating disease progression to augment lacking longitude data. Theoretical analysis also shows that the iterative refining process in PIE (first stage of MedDream) is equivalent to gradient descent with an exponentially decayed learning rate, and practical experiments on three medical imaging datasets demonstrate that MedDream surpasses baseline methods. The clinician A/B test and user study also shows that the disease progression video sequences generated by MedDream are both real and consistent with the corresponding clinical text descriptions. Despite current limitations due to the lack of large amounts of longitude medical imaging data, our framework has vast potential in restoring missing data from previous EHRs, improving clinical education. Moving forward, we aim to incorporate more types of medical imaging data with richer clinical descriptions into medical video generation, enabling our framework to more precise control over disease simulation through text conditioning.

References

Diabetic Retinopathy Detection, howpublished= https://www.kaggle.com/c/diabetic-retinopathy-detection, 2015.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

Nahlah Algethami, Talha Iqbal, and Ihsan Ullah. Generative ai for biomedical video synthesis: a review. *Artificial Intelligence Review*, 58(12):1–50, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023b.

Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, pages 1–13, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024.

Tong Chen, Shuya Yang, Junyi Wang, Long Bai, Hongliang Ren, and Luping Zhou. Surgsora: Decoupled rgbd-flow diffusion model for controllable surgical video generation. *arXiv preprint arXiv:2412.14018*, 2024.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. *arXiv preprint arXiv:2408.14028*, 2024.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Sarah F Cook and Robert R Bies. Disease progression modeling: key concepts and recent developments. *Current pharmacology reports*, 2:221–230, 2016.

Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.

Salman Ul Hassan Dar, Arman Ghanaat, Jannik Kahmann, Isabelle Ayx, Theano Papavassiliu, Stefan O Schoenberg, and Sandy Engelhardt. Investigating data memorization in 3d latent diffusion models for medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63 (11):139–144, 2020.

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

Tianyu Han, Jakob Nikolas Kather, Federico Pedersoli, Markus Zimmermann, Sebastian Keil, Maximilian Schulze-Hagen, Marc Terwoelbeck, Peter Isfort, Christoph Haarburger, Fabian Kiessling, et al. Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nature Machine Intelligence*, pages 1–11, 2022.

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

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

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly

available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Euijin Jung, Miguel Luna, and Sang Hyun Park. Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. *Pattern Recognition*, 133:109061, 2023.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective, 2024. URL https://arxiv.org/abs/2411.02385.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13 (1):7303, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Meryem Mine Kurt, Ümit Mert Çağlar, and Alptekin Temizel. Progressive disease image generation with ordinal-aware diffusion models. *Diagnostics*, 15(20): 2558, 2025.

Nicholas R Kurtansky, Brian M D'Alessandro, Maura C Gillis, Brigid Betz-Stablein, Sara E Cerminara, Rafael Garcia, Marcela Alves Girundi, Elisabeth Victoria Goessinger, Philippe Gottfrois, Pascale Guitera, et al. The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. *Scientific Data*, 11(1):884, 2024.

Daeun Kyung, Junu Kim, Tackeun Kim, and Edward Choi. Towards predicting temporal changes in a patient's chest x-ray images based on electronic health records. *arXiv preprint arXiv:2409.07012*, 2024.

Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1952, 2019.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*, 2023.

Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–240. Springer, 2024a.

Linyuan Li, Jianing Qiu, Anujit Saha, Lin Li, Poyuan Li, Mengxian He, Ziyu Guo, and Wu Yuan. Artificial intelligence for biomedical video generation. *arXiv preprint arXiv:2411.07619*, 2024b.

Linyuan Li, Jianing Qiu, Anujit Saha, Lin Li, Poyuan Li, Mengxian He, Ziyu Guo, and Wu Yuan. Artificial intelligence for biomedical video generation, 2024c. URL https://arxiv.org/abs/2411.07619.

Yihao Liu, Xu Cao, Tingting Chen, Yankai Jiang, Junjie You, Minghua Wu, Xiaosong Wang, Mengling Feng, Yaochu Jin, and Jintai Chen. From screens to scenes: A survey of embodied ai in healthcare. *arXiv preprint arXiv:2501.07468*, 2025.

Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, and James M Rehg. Efficient learning of continuous-time hidden markov models for disease progression. *Advances in neural information processing systems*, 28, 2015.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical imaging 2018: Image processing*, volume 10574, pages 415–420. SPIE, 2018.

Midjourney. Video generation models as world simulators. https://www.midjourney.com/home, 2024. Accessed: 2024-07-30.

Peter G Mikhael, Jeremy Wohlwend, Adam Yala, Ludvig Karstens, Justin Xiang, Angelo K Takigami, Patrick P Bourgouin, PuiYee Chan, Sofiane Mrah, Wael Amayri, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, pages JCO–22, 2023.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1): 12098, 2023.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023.

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

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

Pika. Pika art – home. https://pika.art/home, 2024. Accessed: 2024-07-30.

PixVerse. Pixverse. https://app.pixverse.ai/, 2024. Accessed: 2024-11-10.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Daniele Ravi, Daniel C Alexander, Neil P Oxtoby, and Alzheimer's Disease Neuroimaging Initiative. Degenerative adversarial neuroimage nets: generating images that mimic disease progression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 164–172. Springer, 2019.

Daniele Ravi, Stefano B Blumberg, Silvia Ingala, Frederik Barkhof, Daniel C Alexander, Neil P Oxtoby, Alzheimer's Disease Neuroimaging Initiative, et al. Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia. *Medical Image Analysis*, 75:102257, 2022.

Nathan Raw. Stable diffusion videos. https://github.com/nateraw/stable-diffusion-videos, 2023.

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

Peter Schulam and Raman Arora. Disease trajectory maps. *Advances in neural information processing systems*, 29, 2016.

Kristen A Severson, Lana M Chahine, Luba Smolensky, Kenney Ng, Jianying Hu, and Soumya Ghosh. Personalized input-output hidden markov models for disease progression modeling. In *Machine Learning for Healthcare Conference*, pages 309–330. PMLR, 2020.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34:6216–6228, 2021.

Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674*, 2024a.

Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024b.

Chenyu Tang, Wentian Yi, Edoardo Occhipinti, Yanning Dai, Shuo Gao, and Luigi G Occhipinti. A roadmap for the development of human body digital twins. *Nature Reviews Electrical Engineering*, 1(3): 199–207, 2024.

Mohamad-Hani Temsah, Rakan Nazer, Ibraheem Altamimi, Raniah Aldekhyyel, Amr Jamal, Mohammad Almansour, Fadi Aljamaan, Khalid Alhasan, Abdulkarim A Temsah, Ayman Al-Eyadhy, et al. Openai's sora and google's veo 2 in action: A narrative review of artificial intelligence-driven video generation models transforming healthcare. *Cureus*, 17(1), 2025.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Alexandre Vallée. Envisioning the future of personalized medicine: Role and realities of digital twins. *Journal of Medical Internet Research*, 26:e50204, 2024.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, and Andrew G Lee. Openai's sora in ophthalmology: revolutionary generative ai in eye health. *Eye*, 38 (13):2502–2503, 2024.

Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, and Benyou Wang. Medgen: Unlocking medical video generation by scaling granularly-annotated medical videos. *arXiv preprint arXiv:2507.05675*, 2025.

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

Zhenbin Wang, Lei Zhang, Lituan Wang, Minjuan Zhu, and Zhenwei Zhang. Optical flow representation alignment mamba diffusion model for medical video generation. *arXiv preprint arXiv:2411.01647*, 2024.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, pages 1–11, 2024.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024a.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024b.

Appendix Contents

In this supplementary material, we provide additional details and analyses to complement the main paper. Section A introduces the preliminaries of diffusion models, while section B presents the theoretical analysis for Proposition 1 and Proposition 2. In Section C, we detail the baseline models used in the main experiments. Section D covers the implementation details of MedDream. Additional visualizations, including failure case analysis and disease sequence comparisons across different medical domains, are provided in Section E. Section F discusses the details of two user studies. Finally, Section G includes a brief discussion on the ethics statement.

Appendix A. Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) are a class of generative models that use a diffusion process to transform a simple initial distribution, such as a Gaussian, into the target data distribution. The model assumes that the data points are generated by iteratively applying a diffusion process to a set of latent variables $x_1, \ldots, x_T$ in a sample space χ. At each time step t, Gaussian noise is added to the latent variables x_t, and the variables are then transformed back to the original space using a learned invertible transformation function. This process is repeated for a fixed number of steps to generate a final output. The latent variable models can be expressed in the following form,

$$\begin{aligned} p_\theta(x_0) &= \int p_\theta(x_{0:T}) dx_{1:T}, \\ \textit{where} \quad p_\theta(x_{0:T}) &\prod_{t=1}^{T} p_\theta^{(t)}(x_{t-1}|x_t) \end{aligned} \tag{12}$$

Because of a special property of the forward process,

$$\begin{aligned} q(x_t|x_0) &= \int q(x_{1:t}|x_0) dx_{1:(t-1)} \\ &= \mathcal{N}(x_t; \sqrt{\alpha_t} x_0, (1-\alpha_t) \cdot \mathbf{I}) \end{aligned} \tag{13}$$

we can express x_t as a linear combination of x_0 and a noise variable ϵ, which is the key to enabling the image editing process.

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \epsilon, \quad \textit{where} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{14}$$

Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020a) that uses a non-Markovian forward process to generate data. Unlike Denoising Diffusion Probabilistic Models (DDPM), DDIM does not require explicit modeling of the latent variables. Instead, the model generates samples by solving a non-linear differential equation, which defines a continuous-time evolution of the data distribution. We can express its forward process as follows,

$$q_\sigma(x_{1:T}|x_0) := q_\sigma(x_T|x_0) \prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0) \tag{15}$$

where $q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_T} x_0, (1-\alpha_T), \mathbf{I})$ and for all $t > 1$

$$\begin{aligned} & q_\sigma(x_{t-1}|x_t, x_0) \\ &= \mathcal{N}(\sqrt{\alpha_{t-1}} x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t} x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 \mathbf{I}) \end{aligned} \tag{16}$$

Setting $\sigma_t = 0$, it defines a generation process going from x_t to x_{t-1} as follows

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}} \Big(\frac{x_t - \sqrt{1-\alpha_t} \epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \Big) + \\ \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta^{(t)}(x_t) \end{aligned} \tag{17}$$

where the $\epsilon_\theta^{(t)}(x_t)$ is a model that attempts to predict $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ from x_t

Appendix B. Theoretical Analysis

B.1. Proof of Proposition 1

In this proof, we follow the conventions and definitions in (Song et al., 2020a)

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}} \Big(\frac{x_t - \sqrt{1-\alpha_t} \epsilon_\theta^{(t)}(x_t, y)}{\sqrt{\alpha_t}} \Big) + \\ \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta^{(t)}(x_t, y) \end{aligned} \tag{18}$$

Now given a base image denoted as $\mathbf{x_0^0}$, we wish to perform diffusion-based editing recursively for N times. The roll-back (to the k th-steps, where $k \geq 1$) according to (14):

$$x_n^k = \sqrt{\alpha_k} \cdot x_{n-1}^0 + \sqrt{1-\alpha_k} \cdot \epsilon \tag{19}$$

where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$. Plugging (19) into (18),

$$\begin{aligned} x_n^{k-1} &= \sqrt{\alpha_{k-1}}\left(\frac{x_n^k - \sqrt{1-\alpha_k}\epsilon_\theta(k)(x_n^k, y)}{\sqrt{\alpha_k}}\right) + \\ &\sqrt{1-\alpha_{k-1}} \cdot \epsilon_\theta^{(k)}(x_n^k, y) \\ &= \sqrt{\alpha_{k-1}} \cdot x_{n-1}^0 + \\ &\sqrt{\frac{\alpha_{k-1}(1-\alpha_k)}{\alpha_k}} \cdot (\epsilon - \epsilon_\theta^{(k)}(x_n^k, y)) + \\ &\sqrt{1-\alpha_{k-1}} \cdot \epsilon_\theta^{(k)}(x_n^k, y) \end{aligned} \tag{20}$$

Setting the last step to perform the revert diffusion process

$$\begin{aligned} x_n^0 &= \sqrt{\alpha_0} \cdot x_{n-1}^0 + \\ &\sqrt{\frac{\alpha_0(1-\alpha_1)}{\alpha_1}} \cdot (\epsilon - \epsilon_\theta^{(1)}(x_n^1, y)) + \\ &\sqrt{1-\alpha_0} \cdot \epsilon_\theta^{(1)}(x_n^1, y) \end{aligned} \tag{21}$$

Unrolling this recursion in (21), we have

$$\begin{aligned} x_N^0 &= (\sqrt{\alpha_0})^N \cdot x_0^0 + \sqrt{\frac{\alpha_0(1-\alpha_1)}{\alpha_1}} \cdot \sum_i^N (\sqrt{\alpha_0})^i \cdot \epsilon \\ &+ \frac{\sqrt{\alpha_1 - \alpha_0\alpha_1} - \sqrt{\alpha_0 - \alpha_0\alpha_1}}{\sqrt{\alpha_1}} \cdot \sum_i^N (\sqrt{\alpha_0})^i \cdot \epsilon_\theta^{(i)}(x_i^1, y) \end{aligned} \tag{22}$$

Typically, α_0 is set to a number close to but less than 1, where in the case of stable diffusion 0.9999, α_1 is 0.9995, assuming 50 step schedule.

The second term in (22) is a sampling from Gaussian distributions with geometrically decreasing variances.

$$\lim_{N\to\infty} \sum_i^N (\sqrt{\alpha_0})^i \cdot \epsilon = 0 \tag{23}$$

Given a large enough N,

$$\begin{aligned} x_N^0 &\approx (\sqrt{\alpha_0})^N \cdot x_0^0 + \\ &\frac{\sqrt{\alpha_1 - \alpha_0\alpha_1} - \sqrt{\alpha_0 - \alpha_0\alpha_1}}{\sqrt{\alpha_1}} \cdot \sum_i^N (\sqrt{\alpha_0})^i \cdot \epsilon_\theta^{(i)}(x_i^1, y) \end{aligned} \tag{24}$$

Proposition 1 in (Song et al., 2020a) declares "optimal $\epsilon_\theta^{(i)}$ has an equivalent probability flow ODE corresponding to the "Variance-Exploding" SDE in (Song et al., 2020b)". Hence, $\epsilon_\theta^{(i)}(x_i^1, y) := \nabla_x \log p(x|y)$. This can be seen as gradient descent with a geometrically decaying learning rate with a factor of $\sqrt{\alpha_0}$, with "base learning rate" $\frac{\sqrt{\alpha_1 - \alpha_0\alpha_1} - \sqrt{\alpha_0 - \alpha_0\alpha_1}}{\sqrt{\alpha_1}}$.

Notice that in (24), there is a decaying factor on the initial image x_0^0, as N grows the original image will surely diminish. Therefore, some other empirical measures are required to preserve the structure of the image, such as segmentation masking, edit strength scheduling and etc.

B.2. Additional Theoretical Analysis

B.2.1. Proof of Proposition 2

Proof Continuing on 24, for any $n > 1$, we have

$$\begin{aligned} &\|x_n^0 - x_{n-1}^0\| = \\ &\|(\sqrt{\alpha_0})^n \cdot [(1 - \frac{1}{\sqrt{\alpha_0}}) \cdot x_0^0 - \lambda \cdot \epsilon_\theta^{(n)}(x_n^1, y)]\| \\ &\leq (\sqrt{\alpha_0})^n \cdot [\|(1 - \frac{1}{\sqrt{\alpha_0}}) \cdot x_0^0\| + \lambda \cdot \|\epsilon_\theta^{(n)}(x_n^1, y)\|] \\ &\leq (\sqrt{\alpha_0})^n [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2] \end{aligned} \tag{25}$$

to guarantee $\|x_n^0 - x_{n-1}^0\| \leq \delta$, we just need to set,

$$\begin{aligned} &(\sqrt{\alpha_0})^n [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2] < \delta \\ &\frac{n}{2} \log(\alpha_0) + \log((\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2) < log(\delta) \end{aligned} \tag{26}$$

hence, given that $\alpha_0 < 1$ and $\log(\alpha_0) < 0$

$$n > \frac{2}{\log(\alpha_0)} \cdot (log(\delta) - \log((\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2)) \tag{27}$$

Let $C = \log((\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2)$, we have

$$n > \frac{2}{\log(\alpha_0)} \cdot (log(\delta) - C) \tag{28}$$

From above, we can conclude that as n grows bigger the changes between steps would grow smaller. The difference between steps will get arbitrarily small.

B.2.2. Additional Theoretical Analysis

We use Proposition 3 as an additional support for Proposition 2.

Proposition 3 *For all* $N > 1$, $\|x_N^0 - x_0^0\| \leq [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2]$

Proof From 25 and applying triangle inequality, we observe that the difference is a sum of a geometric sequence scaled by a constant factor,

$$\begin{aligned} \|x_N^0 - x_0^0\| &\leq \sum_{n=1}^{N} \|x_n^0 - x_{n-1}^0\| \\ &\leq \sum_{n=1}^{N} (\sqrt{\alpha_0})^n [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2] \\ &= \frac{1 - (\sqrt{\alpha_0})^N}{1 - \sqrt{\alpha_0}} \cdot [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2] \end{aligned} \tag{29}$$

As N goes to infinity,

$$\begin{aligned} &\lim_{N \to \infty} \|x_N^0 - x_0^0\| \leq \\ &\frac{1}{1 - \sqrt{\alpha_0}} \cdot [(\frac{1}{\sqrt{\alpha_0}} - 1) \cdot C_1 + \lambda \cdot C_2] = \kappa \end{aligned} \tag{30}$$

Proposition 2 and 3 show as n grows bigger, the changes between steps would grow smaller. Eventually, the difference between steps will get arbitrarily small. Hence, the convergence of PIE is guaranteed, and modifications to any inputs are bounded by a constant.

Appendix C. Baselines

Stable Video Diffusion Video (SVD) and Style-Based Manifold Extrapolation (Extrapolation) are two leading techniques in the field of training-free and zero-shot progressive video generation (pretrained model do not have video training data), each displaying promising results within specific domains. However, their applicability remains confined to these particular domains and poses a challenge in extending to the broader scope of different medical imaging data. To illustrate this, Figure 5 in the main body of our paper provides a comparative visualization of multi-step editing using these three techniques.

Stable Video Diffusion Implementation. (SVD) (Raw, 2023) control the multi-step denoising process in the Stable Diffusion Videos pipeline. By smoothly and randomly traversing through the sampled latent space, SVD demonstrates its capability to generate a series of images that progressively align with a given text prompt (see Figure C).

As the state-of-the-art publicly available pipeline, SVD can generate sequential imaging data by interpolating the latent space via multi-step Stable Diffusion. Though SVD is useful for general domain (Raw, 2023), it is not controllable for medical prompts.

Style-Based Manifold Extrapolation Implementation. (Extrapolation) (Han et al., 2022), involves iteratively modifying images by extrapolating between two latent manifolds. To determine the directions of latent extrapolation, the nearest neighbors algorithm is employed on distributions of known trajectories. However, in cases where progression data is not readily available, as in the study at hand, the directions are obtained by randomly sampling and computing the mean of each manifold.

The actual interpolation of Extrapolation for each step can be defined as:

$$\Delta = \frac{1}{m} \sum_{m}^{i=1} \frac{\Delta t^i}{\Delta T} [G^{-1}(x_{n+1}^0) - G^{-1}(x_n^0)] \tag{31}$$

where $G^{-1}(x_n^0)$ is the corresponding latent vector of the image at stage n.

Appendix D. Implementation Details & Reproducibility

In this section, We present the implementation details of the implementation of Transition Generation Process. We also present the system prompt for the GPT-4 re-captioning.

Video Transition Generation Process. To generate transition between each disease state, we used the pretrained weight from LaVie (Wang et al., 2023), SEINE (Chen et al., 2023), and set the $FPS = 16$ in transition generation. ROI masks are also used to control the transition generation as showed in Equation 11.

DeepAUC DenseNet121 Training. We have previously outlined the concept of a classification confidence score. In order to pre-train the DeepAUC

DenseNet121 model, we utilize the original code repository provided by the authors. The model training process involves the use of an exponential learning rate scheduler, with the base learning rate set to 1×10^{-4}. Initially, the model is pre-trained on a multi-class task for each respective dataset. Subsequently, we employ the AUC loss proposed in the study by (Yuan et al., 2021) to finetune the binary classification task, distinguishing between negative (healthy) and positive (disease) samples. The AUC loss finetuning process involves the use of an exponential learning rate scheduler, with the base learning rate set to 1×10^{-2}. All models used for calculating classification confidence score are finetuned for 10 epochs on one NVIDIA L40S GPU, with a batch size of 128. For each class, the final classification accuracy on the validation set is 95.0 % using the finetuned DenseNet121 model.

System Prompt for GPT-4 report re-captioning. In Figure 8, we showed the system prompt of the GPT-4 Re-captioning module in the MedDream.

ROI Mask Generation. The ROI Masks used in experiments are generated by finetuned medical Segment Anything Model (Med-SAM) with medical prior knowledge (Ma et al., 2024; Kirillov et al., 2023) and post-process with edge detection to smooth and enlarge the edge (Canny, 1986). The mask tool is selected based on the input modality and can be detect during GPT-4o recaptioning. For different domains, since the region guide prior is different, the mask shape and size are also different. Figure 9 showcases examples of ROI masks utilized for simulation in the PIE and baseline models.

Time Cost Analysis. Given that both PIE and the baseline methods utilize the same Stable Diffusion backbone (Rombach et al., 2022), a comparison of latency among each method is unnecessary. For simulating disease progression on an image of size 512×512, per step PIE requires approximately 0.078s to generate the subsequent stage when the strength parameter, γ, is set to 0.5, batch size is set to 1 and using one NVIDIA A100 (80GB).

Appendix E. Visualization

To investigate the influence of each hyper-parameter in PIE and analyse the progression visualization, we conduct several ablation studies for visualization, failure cases analysis.

E.1. Visualization for Three Medical Imaging Domains

To provide an in-depth understanding of how PIE performs during different disease progression inference scenarios compared to baseline models, we present detailed visualizations demonstrating PIE's advantages. PIE consistently maintains the realism of the input image even after 10 steps of progression, excelling in most scenarios. Figure 10 displays a comparison among three methods simulating Cardiomegaly progression. PIE outperforms both SVD and Extrapolation by expanding the heart without introducing noise after 2 steps. Figure 11 displays a comparison among three methods simulating Diabetic Retinopathy progression. PIE outperforms both SVD and Extrapolation by adding more bleeding (red) and small blind (white) regions without introducing noise after 2 steps. Figure 12 displays a comparison among three methods simulating Melanocytic Nevi progression. PIE outperforms the other two methods as it keeps the color and shape of the patient's skin but continually enhances the feature for Melanocytic Nevi.

SVD, while it can interpolate within the prompt latent space, tends to generate noise. Although it can generate convincing video sequences in the general domain, it struggles to simulate authentic disease progression. On the other hand, Extrapolation, despite being effective for bone X-rays, faces challenges with more complex medical imaging domains like chest X-rays. Extrapolation's editing process is considerably slower than the other two methods, and it fails to be controlled effectively by clinical reports.

E.2. Failure Case Analysis

Despite outperforming baseline models, PIE still faces limitations tied to data sensitivity issues. For instance, imbalances in the distribution of training data for Stable Diffusion can limit PIE's capability to edit rare diseases. In some cases, PIE might generate essential medical device features but fail to preserve them in subsequent stages of progression simulation, as observed with features like pacemakers. Figure 13 shows a good example of pacemaker disappearance during simulation. Besides, the co-occurring diseases simulation also has some failure cases (see Figure 15).

Fundamentally, these shortcomings could be addressed with a larger and label-equal distribution dataset. However, given that the volume of medical data is often smaller than in other domains, it is also important to explore fine-tuning PIE's diffusion

Prompt template for GPT-4 Re-captioning

System Message
You are a helpful assistant that generate clear and short clinical report.
I will give you raw clinical text input: ...
The definition of diseases with chest X-ray, diabetic retinopathy, and benign skin lesion.
An example of raw clinical report: ...
An example of simplified clinical report: ...
You should only respond in the format as described below:

Response Format
Report: The medical report for disease progression generation.

Figure 8: Prompt template for system message and response format for GPT-4 clinical report re-captioning.

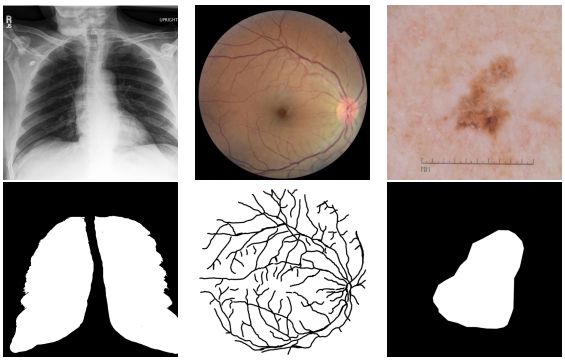

Figure 9: ROI masks for three different domains. Note, the white part in the mask is the disease-related regions.

backbone through few-shot learning under extremely imbalanced label distribution.

E.3. Additional Discussion for Hyperparameter Search & Analysis

To demonstrate the significance of the ROI mask as a key control factor in the PIE, an experiment was conducted to compare the performance of models using and not using ROI masks across three domains, with each model tested using 5 different random seeds. The evaluation focused on the classification confidence score and CLIP score. Further visual analysis depicted in Figure 16 shows that the ROI mask crucially helps in preserving the basic shape of the medical imaging during the PIE process. Consequently, these findings suggest that the ROI mask, alongside clinical reports, serves as a critical medical prior for simulating disease progression. It helps the PIE to concentrate on disease-related regions while maintaining the realism of the input image.

We present a comprehensive examination of various hyperparameters, namely strength, progression stage (N), β_1, and β_2, and their respective tradeoffs as demonstrated in tables 7, 8, and 9. Notably, a discernible tradeoff exists between classification confidence score and CLIP-I/KID, wherein an increase in classification confidence score results in a decrease in CLIP-I.

Table 7 illustrates a positive correlation between the strength of PIE and classification confidence score, while revealing a negative relationship between strength and CLIP-I/KID. From an intuitive standpoint, as the strength of PIE increases, more features are directed to align with the pathologies within the original images. Consequently, the classifier exhibits greater confidence in accurately predicting the specific disease class, leading to a more significant deviation from the initial starting point. As a result, the classification confidence score value increases while the CLIP-I value decreases, reflecting the inverse relationship between the two metrics.

Table 8 presents a similar pattern initially. However, both classification confidence score and CLIP-I/KID reach a state of stability after a certain number of steps. This observation aligns with our theoretical analysis as outlined in Proposition 2. Moreover, it can be interpreted as the convergence of a Cauchy geometric sequence, wherein the discrepancy between successive steps gradually diminishes as the value of N tends towards infinity.

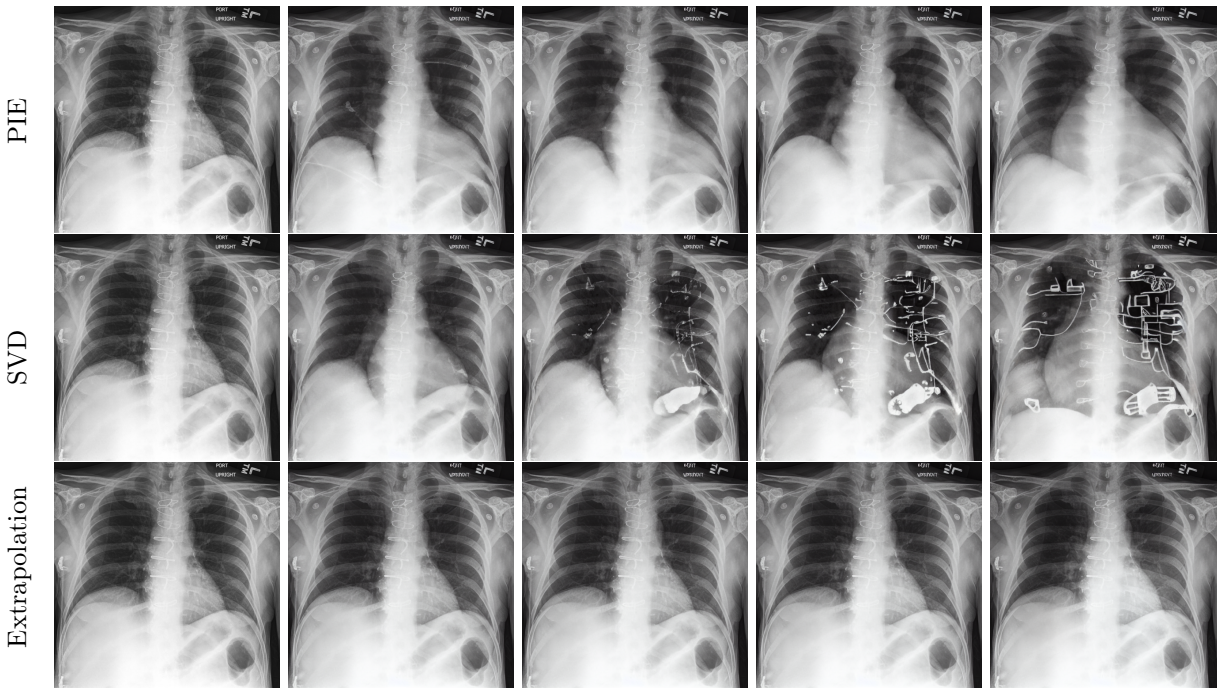

Figure 10: Visualization of PIE, SVD, Extrapolation to generate disease progression from Cardiomegaly clinical reports.

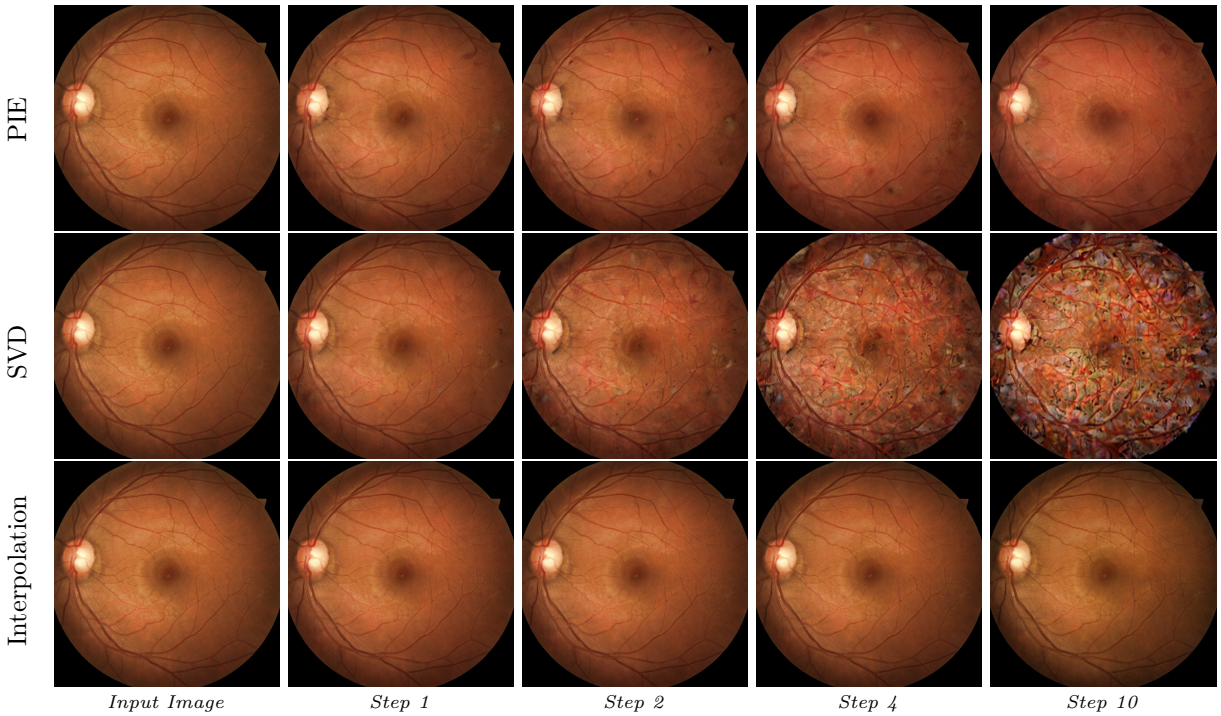

Figure 11: Visualization of PIE, SVD, Extrapolation to generate disease progression from Diabetic Retinopathy clinical reports.

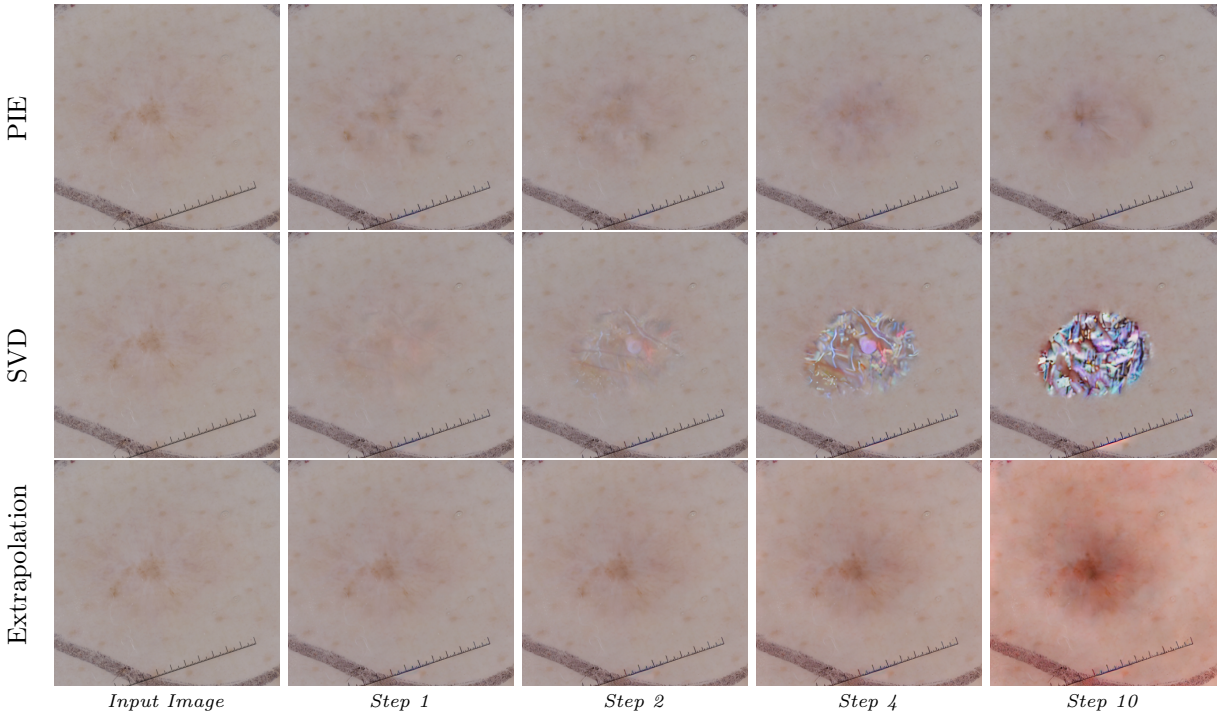

Figure 12: Visualization of PIE, SVD, Extrapolation to generate disease progression from Melanocytic Nevi clinical reports.

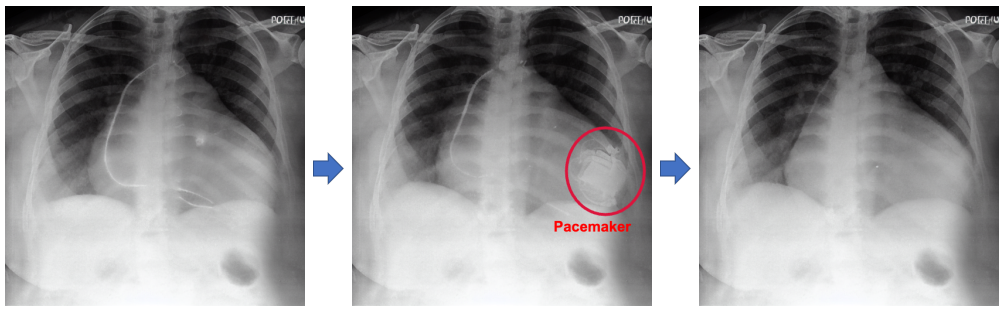

Figure 13: A failure case of the PIE model in preserving the features of a pacemaker during the simulation of Cardiomegaly disease progression. Pacemaker is usually used by patients with severe Cardiomegaly. At Step $n-1$, the X-ray displays an electronic line. At Step n, both the electronic line and the pacemaker are visible. However, by Step $n+1$, all the medical device features, including the pacemaker, have vanished from the simulation. It's important to note that the input clinical prompt did not contain any information regarding the pacemaker, making it difficult for the model to retain this crucial feature. This illustrates the challenges faced by models like PIE in dealing with significant but unmentioned clinical features in the input data. It underscores the need for incorporating comprehensive and detailed clinical data to ensure accurate and realistic disease progression simulations.

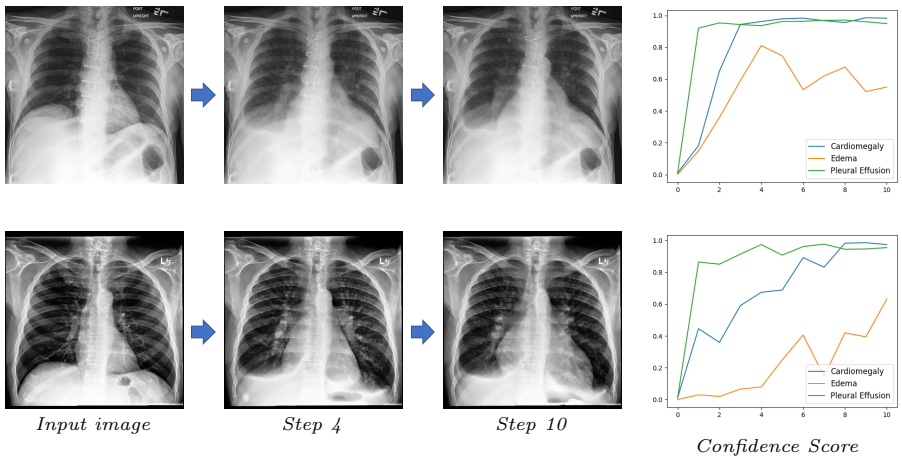

Figure 14: PIE can successfully simulate co-occurring disease progression (Patent's clinical report shows high probability to be Cardiomegaly, Edema, Pleural Effusion at the same time).

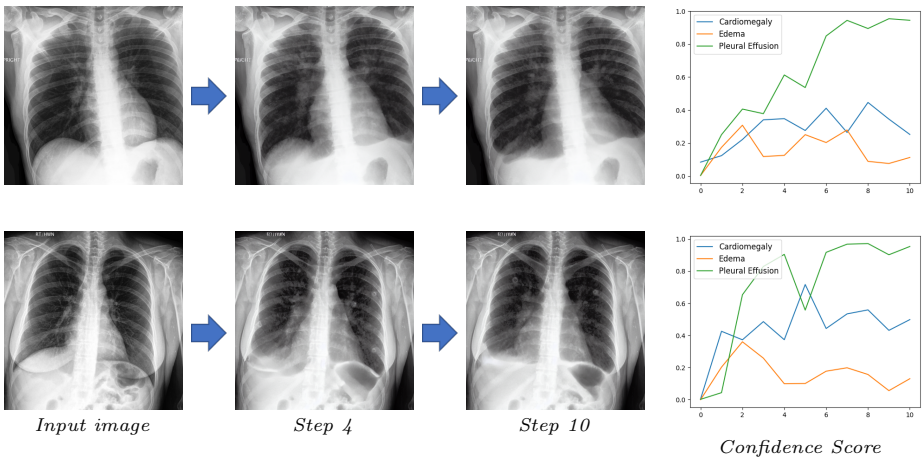

Figure 15: Two failure cases of the PIE model in simulating co-occurring diseases progression for Cardiomegaly, Edema, and Pleural Effusion, only the features for Pleural Effusion are captured. These failure cases arise from the issue related to imbalanced label distribution in the training data. Specifically, the prevalence of Pleural Effusion is significantly higher than the other four classes, leading to an inherent bias in the model's simulations for co-occurring diseases. This imbalance emphasizes the need for a more diversified and balanced training dataset for more accurate simulation of co-occurring diseases.

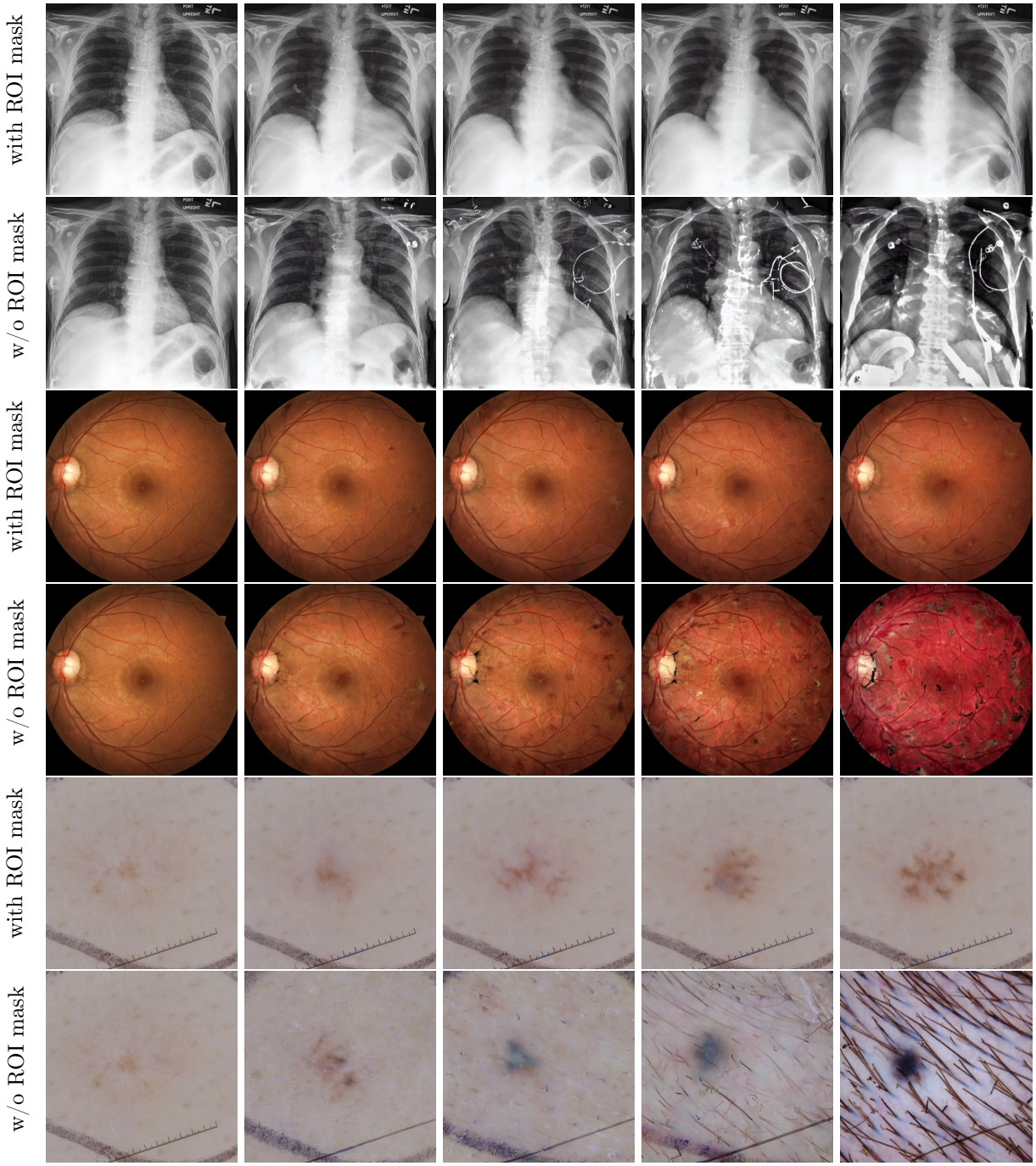

Figure 16: Visualization comparison between ROI mask influence for three medical imaging domains.

Lastly, in Table 9, we explore the interplay between β_1 and β_2, which serve as parameters governing the rate of progression within and outside the region of interest (ROI) respectively. Our findings reveal that β_2, responsible for regulating the pace of progression within the ROI, exerts a more pronounced influence on classification confidence score, while β_1 exhibits a stronger impact on CLIP-I/KID. This outcome can be intuitively comprehended, considering that the ROI typically encompasses a smaller area of paramount importance for aligning with disease-specific features. Conversely, the areas outside the ROI exert a significantly greater influence on the realism captured by CLIP-I/KID.

In summary, our study elucidates the inherent trade-offs and offers valuable practical insights, thereby furnishing meaningful guidance for effectively utilizing PIE in practice.

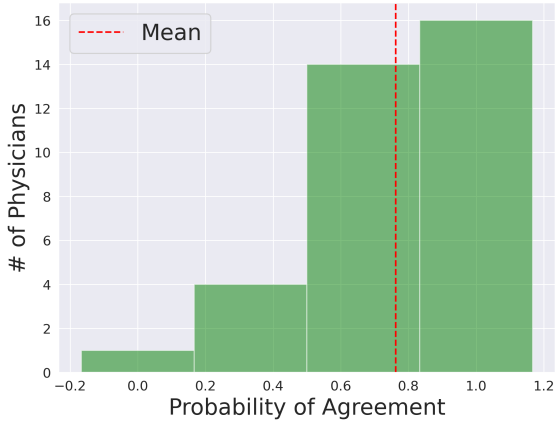

Figure 17: Distribution of probability of agreement to the generated progressions among 35 veteran physicians.

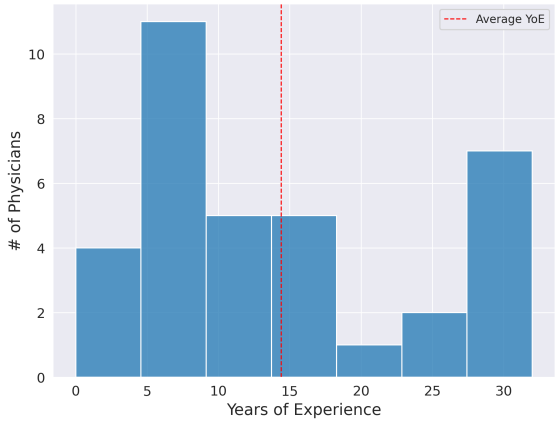

Figure 18: The distribution of the years of experience from the group of physicians who participated in the user study. The average number of years of experience is 14.4 years.

Appendix F. User Study

In our main experiments, we conducted two user studies. The first one is the A/B test from 30 30 clinicians and radiologists as user preference study in section 5.4. The second one is the user feedback study in section 5.6.

F.1. A/B Test for Video Comparison

In the first user study, we conduct an user study from 24 medial doctors and 6 medical students. Participants came from different departments, including Department of Radiology, Department of Dermatology, Department of Ophthalmology and Department of Pediatrics. In this user study, each participant will check 18 pairs of video randomly from MedDream and three baselines (PixVerse (PixVerse, 2024), CogVideoX (Yang et al., 2024), and Pika (Pika, 2024)). For all methods, the input is an image of the initial state of the disease and a prompt to control the progression of the disease. The output is a 3-4s video. In the A/B test, the participant will compare two videos and decide which video is high-fidelity and highly related to the disease progression in the real-world.

F.2. Questionnaire in the User Feedback Study

The questionnaire in the second survey is approved by Affiliated clinical research institute. The questionnaire includes 2 parts. Part one consists of 20 multiple choices of single image classifications, 10 single-step generations, and 10 real X-ray images sampled from the training set. Part two consists of 3 generated disease progressions consisting of Cardiomegaly, Edema as well Pleural Effusion. Each progression runs 10 steps. For each single image classification, we ask "Please determine the pathologies of the following patient" and let the user pick from 6 options {No findings, Cardiomegaly, Consolidation, Edema, Effusion, Atelectasis} with possible co-occurrence, while for each of the 10-step progressions, we ask "Does the

below disease progression fit your expectation?" and let the user input a binary answer of yes or no.

Below we include the full instructions we gave for our user study both in English and Chinese. Here we only include the English version of questions. The examples are shown in Figure 19 and 20.

1. Please read the instructions and inspect the images carefully before answering.
2. Please provide your years of experience
3. For the first 20 questions, please determine the pathologies from the X-ray images (you can choose more than one answer). For the last 3 questions, please answer if the disease progression shown fits your expectations.

F.3. Statistics

The distribution of the years of experience from the group of physicians who participated in the user study. The average number of years of experience is 14.4 years. Over half of the group have more than 10 years of experience. This data attests that our surveyees are highly professional and experienced.

To show the significance of our findings, we also performed the paired t-test on the F1 scores of real and generated scans over the 35 users. Our finding is significant with a p-value of 0.0038.

To quantitatively analyze their responses, we treat each class of pathologies as an independent class and compute precision, recall, and F1 over all the pathologies and physicians.

Appendix G. Discussion and Ethics Statement

The proposed framework is subject to several limitations, with one of the primary constraints being the limited scope of Stable Diffusion v1.4. Due to the model's pre-training on general domain data, the absence of detailed medical reports poses a significant challenge to the model's ability to accurately and reliably edit medical images based on precise text conditioning. Moreover, the framework's overall performance may be influenced by the quality and quantity of available data, which can limit the model's accuracy and generalizability. Moving forward, it would be beneficial to explore ways of integrating more detailed descriptions of medical data in the fine-tuning process of Stable Diffusion to improve the framework's performance and precision in disease simulation through text conditioning. Additional details on the framework's limitations, including an analysis of failure cases, can be found in Supplementary E.

MedDream holds promise as a technology for simulating disease progression, but it also raises concerns regarding potential negative social impacts. One crucial concern revolves around the ethical use of medical imaging data, which may give rise to privacy and security issues. To address this, healthcare providers must take measures to safeguard patient privacy and data security when utilizing MedDream. An effective mitigation strategy involves employing anonymized or de-identified medical imaging data, while also adhering to ethical guidelines and regulations like those outlined by HIPAA (Health Insurance Portability and Accountability Act).

Another concern relates to the accuracy of the simulations generated by the framework, as errors could lead to misdiagnosis or incorrect treatment decisions. To alleviate this concern, rigorous testing and evaluation of the technology must be conducted before its implementation in a clinical setting. Enhancing the accuracy of simulations can be achieved by incorporating additional data sources, such as patient history, clinical notes, and laboratory test results. Additionally, healthcare providers should receive adequate training on effectively utilizing the technology and interpreting its results.

Discrimination against certain groups, based on factors such as race, gender, or age, poses yet another potential concern. Healthcare providers must ensure the fair and unbiased use of the technology. This can be accomplished by integrating diversity and inclusion considerations into the technology's development and training processes, as well as regular monitoring and auditing its usage to identify any signs of bias or discrimination.

The cost and accessibility of the technology present further concerns, potentially restricting its availability to specific groups or geographic regions. To tackle this issue, healthcare providers should strive to make the technology accessible and affordable to all patients, regardless of socioeconomic status or geographic location. This can be achieved through the creation of cost-effective models, partnerships with healthcare providers, and government funding initiatives.

Lastly, there is a risk of excessive reliance on technology, leading to a diminished reliance on clinical judgment and expertise. To mitigate this concern,

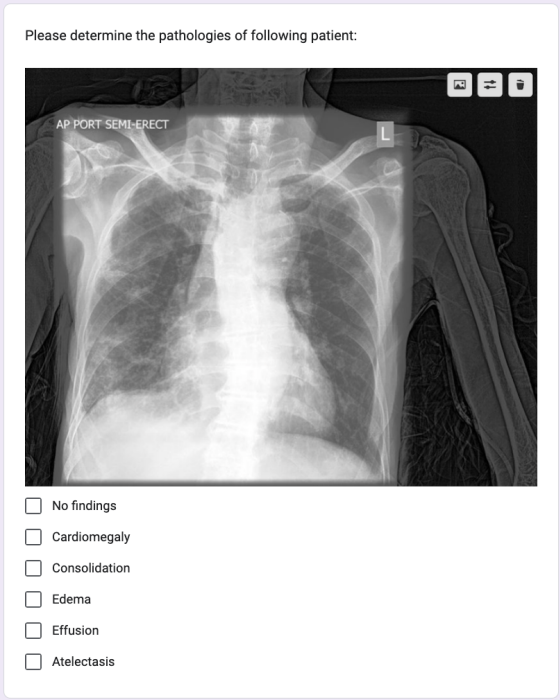

Figure 19: Example of User Study I: we ask the physicians to pick from the 6 options and it's possible to pick more than one option.

Figure 20: Example of User Study II: we ask the physicians to decide if the generated progression of the disease is credible or not.

healthcare providers must be trained to view the technology as a tool that complements their clinical judgment and expertise, rather than relying solely on it for diagnostic or treatment decisions. The technology should be used in conjunction with other data sources and clinical expertise to ensure a comprehensive understanding of disease progression.

In conclusion, while the use of MedDream comes with potential negative social impacts, there are viable mitigations that can address these concerns. Healthcare providers must be aware of the ethical implications associated with this technology and take appropriate measures to ensure its safe and responsible utilization.