Indian Institute of Technology Kanpur

Course Project

Supervisor - Prof. Arnab Bhattacharya (CSE) and Prof. Subhajit Roy (CSE)

Department of Computer Science Engineering

Self Improving Generative Foundation Model for Synthetic Medical Image Generation

Dhruv Mittal (220363)

Harsh Pratap Singh (220432)

Ishan Dandwani (220461)

2025-2026

Contents

1 Introduction

Recent advances in generative AI have enabled synthetic image generation with unprecedented realism, yet producing clinically meaningful medical images remains a significant challenge. Unlike natural images, medical imaging requires strict anatomical accuracy, modality-specific structures, and high diagnostic fidelity—attributes essential for clinical decision-making. Standard text-to-image models such as Stable Diffusion, although powerful in general visual domains, are not inherently optimized for medical interpretation and therefore struggle to generate images that satisfy these stringent diagnostic requirements. Additionally, high-quality medical datasets are scarce, expert annotations are expensive, and privacy regulations such as HIPAA and GDPR severely restrict the direct sharing of patient data. These limitations create a pressing need for systems capable of synthesizing reliable, privacy-preserving, and diagnostically useful medical images.

In this context, synthetic medical image generation has emerged as a promising direction for healthcare AI. Synthetic data can augment limited datasets, reduce annotation burdens, support training of downstream diagnostic models, and provide privacy-preserving alternatives to real patient data. However, achieving this in a clinically valid way requires generative models that not only produce visually plausible images but also encode anatomical and pathological correctness.

This project presents a comprehensive diffusion-based text-to-medical-image generation system enhanced by a two-stage Reinforcement Learning from Human Feedback (RLHF) training pipeline. The approach integrates domain-adapted Stable Diffusion with medical-specific text conditioning, expert-driven quality assessment, and automated alignment models to generate high-fidelity synthetic medical images. While diffusion models provide the generative foundation, RLHF guides the system toward producing images that are not just visually realistic but also clinically meaningful.

Our work focuses specifically on the generation of fundus images, which play a critical role in diagnosing ophthalmic diseases such as diabetic retinopathy, glaucoma, hypertensive retinopathy, and macular degeneration. Fundus imaging presents unique challenges due to its highly structured appearance—optic disc, macula, vessel architecture, retinal pigmentation—and its sensitivity to subtle pathological variations. Generating synthetic fundus images that accurately reflect these structures requires both domain-specific fine-tuning and medically informed training strategies.

2 Background Knowledge

2.1 Text-to-Image Generation

Text-to-image generation aims to synthesize an image x from a descriptive text prompt c by learning a conditional distribution $p(x \mid c)$. The model must understand semantic information in the text and produce a visually coherent image that matches the described content. In medical imaging, this requires not only realism but also anatomical correctness and modality-specific structures.

2.1.1 Previous Approaches

Earlier methods relied heavily on Generative Adversarial Networks (GANs), such as StackGAN and AttnGAN, which attempted to map text embeddings to pixel space. Although capable of generating natural images, GANs suffered from common issues including:

- instability during adversarial training,
- mode collapse and limited diversity,
- difficulty preserving fine anatomical details,
- and poor alignment between text and image for medical prompts.

Some hybrid approaches introduced attention mechanisms to improve text–image consistency, but they still lacked the stability and resolution needed for high-quality medical image synthesis.

2.1.2 Why Diffusion Models Work Better

Diffusion models overcome these limitations by learning a stable denoising process that reconstructs images from noise. When combined with transformer-based text encoders such as BioMedBERT, the model can accurately interpret medical terminology (e.g., "optic disc", "vessel dilation") and guide image synthesis through cross-attention.

Operating in latent space via a VAE further reduces computation while preserving clinical structure. In our system, diffusion is additionally enhanced by a two-stage RLHF pipeline where a Selector Model filters low-quality outputs and a Policy Model improves alignment. This leads to:

- significantly better text–image consistency,
- higher anatomical fidelity in fundus structures,
- more stable and diverse generations,
- and improved clinical realism compared to GAN-based methods.

Overall, diffusion models combined with medical-domain text encoders and RLHF provide a more robust, accurate, and medically meaningful approach to text-driven fundus image synthesis.

2.2 Diffusion

A diffusion model is a generative model that *learns to make images by gradually removing noise*. It consists of two complementary processes:

1. Forward diffusion (corruption): starting from a real image x_0, small Gaussian noise is added repeatedly to produce a sequence

$$x_1, x_2, \ldots, x_T,$$

until x_T is nearly pure noise. This forward process is typically defined by

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\big),$$

where $\{\beta_t\}$ is a predefined noise schedule.

2. Reverse diffusion (generation): the model learns a reverse denoising process $p_\theta(x_{t-1} \mid x_t)$ that removes noise step-by-step, starting from random noise x_T and producing a clean image x_0. In practice the model predicts the noise component $\varepsilon_\theta(x_t, t)$ and uses it to recover x_{t-1}.

Why it works:

- Training is stable (predicting noise is an easy supervised target).
- Produces high-quality, detailed images.
- Can be *conditioned* on text (or other inputs): the denoiser is given a text embedding so the reverse process generates images matching the prompt.

In one sentence: diffusion models learn how images are corrupted by noise and then invert that process to generate new, realistic images from pure noise.

2.3 RLHF

Reinforcement Learning from Human Feedback (RLHF) is a training method where a model improves its outputs by using feedback that comes from humans (or human-trained evaluators). Instead of learning only from fixed datasets, the model learns what *humans consider high-quality.*

How it works in simple steps:

1. **Generate:** the model creates several candidate images from a text prompt.
2. **Evaluate:** a human-trained *Selector Model* scores these images based on quality and correctness.
3. **Reward:** good images get high scores (reward), bad images get low scores (penalty).
4. **Learn:** a *Policy Model* updates itself to produce images that receive higher rewards.

Why RLHF is useful:

- Ensures generated images match human expectations.
- Encourages clinical correctness for medical image synthesis.
- Allows the system to "self-improve" over time.

In our project:

- The **Selector Model** checks fundus image quality (optic disc clarity, vessel correctness, artifacts).
- The **Policy Model** adjusts the generator to produce more accurate and medically meaningful images.
- This two-stage RLHF loop improves the final fundus images beyond what diffusion alone can achieve.

In one sentence: RLHF teaches the model to generate better medical images by learning from human-influenced quality scores.

2.4 Transformers and Variational Autoencoders (VAE)

Modern text-to-image systems rely heavily on two key components: transformer-based text encoders and Variational Autoencoders (VAEs). Together, they allow the model to understand complex textual descriptions and synthesize images efficiently in a compressed latent space.

2.4.1 Transformers for Text Understanding

Transformers use self-attention to capture long-range dependencies and contextual meaning within a text prompt. For medical image synthesis, transformer models such as BioMedBERT provide a major advantage because they are trained on biomedical literature and can accurately represent medical terms (e.g., "optic disc edema", "microaneurysms"). This creates a rich embedding E that conditions the generative model through cross-attention mechanisms.

2.4.2 VAE for Latent Image Representation

A Variational Autoencoder compresses an input image into a latent vector z using an encoder and reconstructs it using a decoder. Instead of operating in pixel space, diffusion models perform all denoising steps in this latent space:

$$z = \mathrm{VAE}_{enc}(x), \qquad x = \mathrm{VAE}_{dec}(z).$$

This reduces computational cost while preserving important retinal structures such as the macula, optic disc, and vessel patterns.

3 Architecture

The proposed system combines a diffusion-based text-to-fundus generator with a two-stage Reinforcement Learning from Human Feedback (RLHF) refinement pipeline. The architecture is designed to understand medical text accurately, synthesize anatomically consistent fundus images, and iteratively improve output quality based on learned human feedback.

3.1 Diffusion Model Architecture

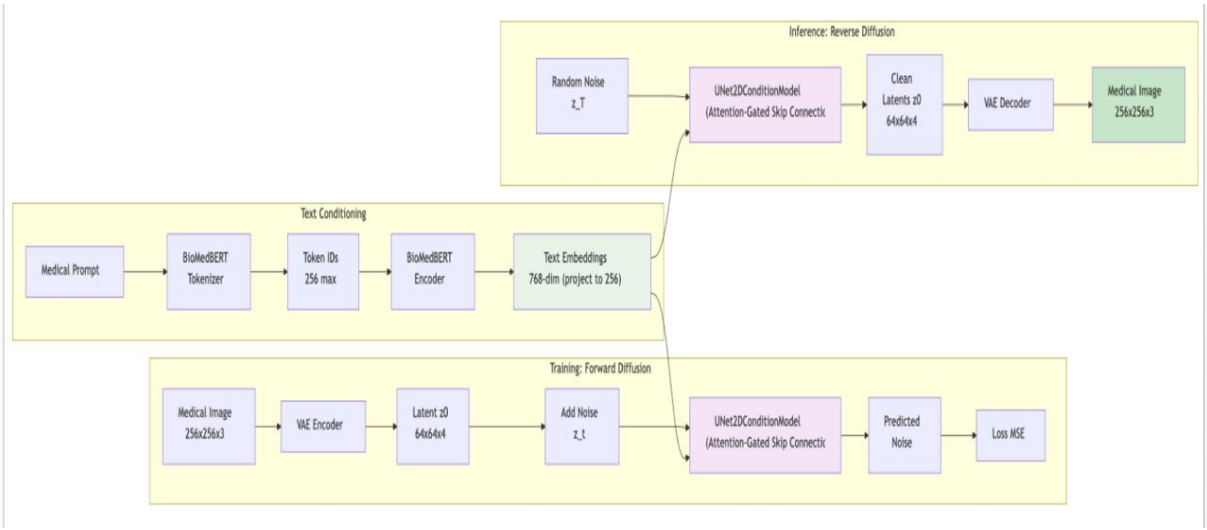

Figure 1: Diffusion-based text-to-image generation pipeline.

The diffusion system operates in three major components: text conditioning, forward diffusion, and reverse denoising.

Text Conditioning

A medical prompt is first tokenized using BioMedBERT and converted into contextual embeddings. BioMedBERT is trained on biomedical literature, enabling accurate interpretation of ophthalmic terms such as *optic disc*, *fovea*, or *retinal vessels*. These embeddings condition the diffusion UNet via cross-attention layers.

Forward Diffusion

For training, real fundus images are encoded into latent space using a VAE. Controlled Gaussian noise is added over a sequence of timesteps, generating progressively degraded

latents. This teaches the UNet how images become noisy, forming the basis for the reverse process.

Reverse Diffusion (Inference)

During generation, sampling begins from pure random noise. The UNet denoises this latent step-by-step, guided by the BioMedBERT embeddings. After reaching a clean latent representation, the VAE decoder converts it into a full-resolution fundus image. This process ensures structure preservation and consistent anatomical realism.

3.2 RLHF Training Pipeline

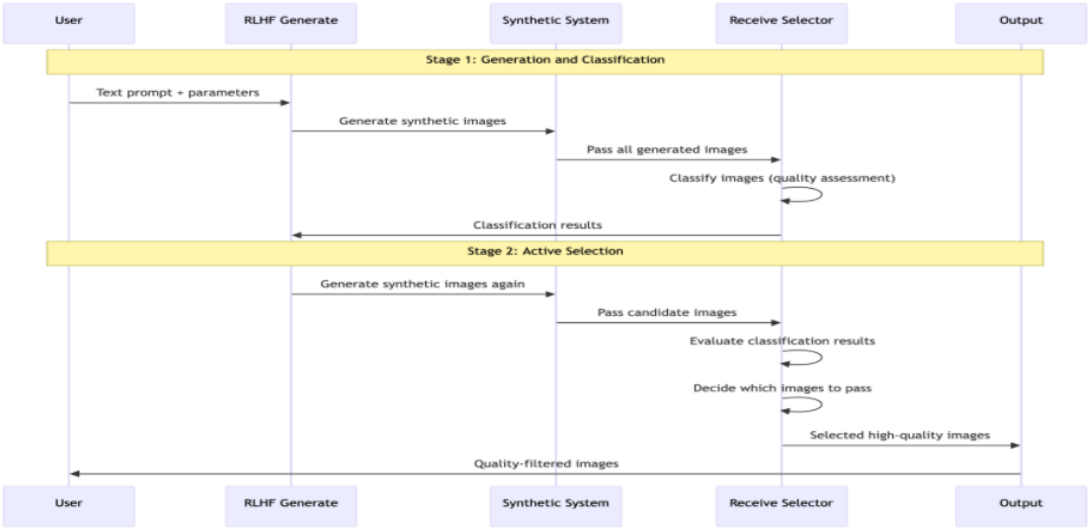

Figure 2: Two-stage RLHF pipeline: classification and active selection.

The RLHF component further improves generation quality by incorporating expert-like feedback.

Stage 1: Generation and Classification

The Synthetic System produces multiple candidate fundus images from a prompt. These images are passed to the Selector Model, which assigns a quality score based on criteria such as optic disc clarity, vessel structure, contrast, and artifact presence. The classification results guide early policy updates.

Stage 2: Active Selection

The generator produces another batch of images, which the Selector scores again. Only high-quality images surpassing a learned threshold are passed forward. The Policy Model then learns a reward signal from these high-scoring samples, adjusting the generator to prefer outputs that resemble clinically correct structures.

This two-stage pipeline stabilizes training and provides consistent improvements in realism and medical correctness.

3.3 Overall Workflow

The complete system follows this sequence:

1. **Input:** A medical text description of a fundus image.
2. **Encoding:** BioMedBERT converts text into medical-aware embeddings.
3. **Generation:** Latent diffusion produces a synthetic fundus image.
4. **Scoring:** The Selector evaluates image quality and correctness.
5. **Reward:** High-scoring samples reinforce the Policy Model.
6. **Refinement:** The generator iteratively improves using RLHF signals.

3.4 Advantages of the Architecture

The combination of diffusion, latent-space modeling, medical-domain transformers, and RL-based quality optimization allows the system to:

- maintain strong text-to-image alignment,
- produce realistic fundus structures,
- reduce artifacts and inconsistencies,
- adapt to expert preferences through feedback,
- and generate diverse, clinically meaningful images.

This architecture provides a reliable and scalable framework for synthetic medical image generation.

3.5 System Architecture Overview

The system consists of five major modules: **Data Input**, **Training Pipeline**, **RLHF System**, **Application Layer**, and **Evaluation**. Together, these components enable fine-tuning of a diffusion model using expert-rated data, iterative reinforcement learning with

human feedback, and final high-quality image generation.

3.5.1 Data Input

The pipeline begins with the collection and preprocessing of structured text–image data:

- **CSV Input:** The dataset contains image file paths, associated textual descriptions, and modality or category labels.
- **Example Data Extraction:** This information is parsed to produce a standardized training dataset for the model.
- **Expert Ratings:** Domain experts provide quality or correctness scores for selected images.
- **Rating System:** Expert evaluations are normalized into a consistent scoring framework used for training the selector (reward) model.

3.5.2 Training Pipeline

The prepared data is used to train and adapt the diffusion model:

- **Model Training:** The system fine-tunes Stable Diffusion on paired text–image inputs to specialize it for the target domain.
- **Fine-tuned Model:** This produces a domain-adapted version of Stable Diffusion capable of generating relevant images.
- **Checkpoint Saving:** Intermediate and final model weights are stored for inference, debugging, or further refinement.

3.5.3 RLHF System

After fine-tuning, the model undergoes an active improvement cycle inspired by Reinforcement Learning with Human Feedback (RLHF):

- **Synthetic Image Generation:** The fine-tuned model generates new synthetic samples.
- **Generated Image Pool:** These samples form a candidate set for evaluation.
- **Selector Model:** A reward model trained on expert ratings predicts the quality and suitability of each synthetic image.
- **Quality Filtering:** Only high-quality and domain-appropriate images are selected.
- **Policy Model Update:** The filtered images are used to reinforce and update the policy model. This creates a feedback loop where the model continually improves

by generating, evaluating, and learning from synthetic data.

This iterative cycle enables the model to progressively align with expert-defined quality standards without requiring continuous manual annotation.

RLHF Training Algorithm Details. The RLHF module is optimized using a REINFORCE-style policy gradient with a two-stage setup. Instead of sampling discrete actions, the policy outputs a continuous alignment score in the range $[0, 1]$, representing how well the generated image matches the input prompt.

- **Continuous rewards:** The model uses normalized human or selector ratings in the range 1–3 as continuous reward signals rather than discrete categorical labels.
- **Implicit action:** The policy does not sample from a set of actions; it directly predicts an alignment score which serves as a continuous control signal.
- **Value function:** In Stage 2, the function `compute_value()` uses the policy's output as a baseline to reduce variance in the policy gradient update.
- **Two-stage training:**
  - **Stage 1:** The model is trained on all generated samples to learn initial reward–policy relationships.
  - **Stage 2:** The Selector filters low-quality samples; the policy is then trained only on high-quality data, emphasizing medically correct structures.

This two-stage REINFORCE setup stabilizes training, provides smoother gradient updates, and aligns the policy with clinically relevant image characteristics.

Selector Model The Selector uses a Vision Transformer to evaluate fundus image quality.

Patch-based Processing

- Splits each fundus image into 16×16 patches.
- Each patch captures local retinal structures (vessels, optic disc, background).
- Every patch is embedded into a 768-dimensional feature vector.

Multi-head Self-Attention

- 12 attention heads over 6 transformer layers.
- Captures relationships between spatial patches.
- Detects:
  - Optic disc clarity (central disc patches).

  - Vessel correctness (branching and continuity).
  - Artifacts (noise or inconsistent patterns).

Learning from Expert Ratings

- Expert annotations: 1–3 scale, normalized to $[0, 1]$.
- Learns via advantage-weighted policy gradient.
- Stage 2 applies a value baseline to reduce variance.

Policy Model The Policy model aligns images and medical prompts to guide downstream generation.

Image Encoder (Vision Transformer)

- Produces 768-dimensional embeddings.
- Encodes optic disc, vessels, and macula features.

Text Encoder (BioMedBERT)

- Encodes medical prompts (e.g., clinical fundus descriptions).
- Over 28,000 medical-domain tokens.
- Maximum length of 256 tokens.

Fusion Network

- Concatenates image (768-dim) and text (768-dim) embeddings into a 1536-dim vector.
- Multi-layer MLP: $1536 \to 768 \to 384$.
- Learns cross-modal alignment.

Alignment Head

- Outputs an alignment score in the range $[0, 1]$.
- Higher values indicate a better image–prompt match.

Training via REINFORCE

- The policy is optimized using the REINFORCE algorithm.
- Learns directly from expert-provided alignment feedback.
- Used in a two-stage selection and scoring pipeline.

3.5.4 Application Layer

Once training and RLHF refinement are completed, the model is used for practical generation tasks:

- **Script Generation:** Model checkpoints are used to generate structured prompts or scripts that drive the generation workflow.
- **Image Generation Engine:** These prompts are fed into the model to generate high-resolution images.
- **Output Images:** The final generated images are presented to the user or downstream applications.

3.5.5 Evaluation

The system includes a quantitative evaluation module for assessing model performance:

- **Metrics Calculation:** The generated outputs are evaluated using standardized objective metrics.
- **FID, IS, and SSIM Scores:**
  - *Fréchet Inception Distance (FID):* Measures how closely the generated distribution matches real data.
  - *Inception Score (IS):* Evaluates image diversity and recognizability.
  - *Structural Similarity Index (SSIM):* Computes perceptual similarity between image pairs.

These metrics collectively provide a reliable assessment of the improvements achieved through fine-tuning and RLHF.

4 Results

Metric	Ground Truth	Fundus 512
SNR	2.3633	2.7365
Contrast	0.6780	0.8065
Sharpness	0.0184	0.1146
Vessel Visibility	0.0056	0.0297
Uniformity	0.1020	0.0588

Table 1: Comparison of image quality metrics between ground truth and Fundus 512.

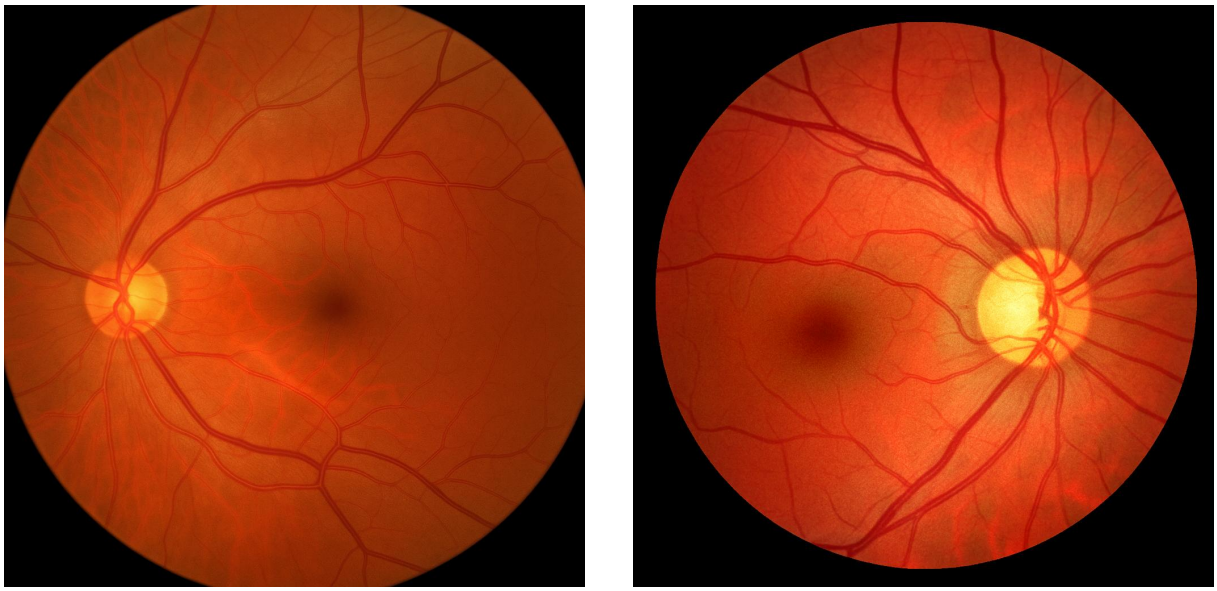

Figure 3: 512px generated fundus images

References

[1] Wang, J., Wang, K., Yu, Y., *et al.*, "Self-improving generative foundation model for synthetic medical image generation and clinical applications," *Nature Medicine*, vol. 31, no. 2, pp. 609–617, 2025.

[2] WithStomach, "MINIM: Medical Image Normalization and Integration Model," GitHub Repository, 2024. Available: https://github.com/WithStomach/MINIM

[3] Gu, Y., Tinn, R., Cheng, H., *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 1, Article 1, pp. 1–24, 2021. DOI: 10.1145/3458754.