Enhancing Empathetic Interaction: A DreamBooth-Based Approach for Personalized Visual Therapy

Shweta Sharma
Department of Computer Science
Drexel University
Philadelphia, PA, USA
ss5792@drexel.edu

Khushboo Patel
Department of Computer Science
Drexel University
Philadelphia, PA, USA
kp3329@drexel.edu

Milan Varghese
Department of Computer Science
Drexel University
Philadelphia, PA, USA
mv644@drexel.edu

***Abstract*—Advances in generative modeling have opened new possibilities in personalized therapeutic applications, where customized visual stimuli can support emotional well-being. In this paper, we present a novel DreamBooth-based fine-tuning approach utilizing a pre-trained Stable Diffusion model to generate personalized visuals from simple text prompts. Our method specifically focuses on context-aware image generation using a dataset comprising personal pet images, thus enabling more emotionally resonant outputs for visual therapy applications. Quantitative evaluation using CLIPScore demonstrates high semantic alignment (average CLIPScore: 83.9%), and qualitative assessments confirm visual coherence and personalization. Despite promising results, we observe limitations in generalization due to dataset size constraints, suggesting directions for future improvement. Our integrated approach highlights the potential and current challenges in employing diffusion-based generative models for empathetic and personalized visual therapies.**

***Index Terms*—Stable Diffusion, DreamBooth, CLIPScore, Image Synthesis, Visual Therapy.**

I. Introduction

Generative modeling has recently witnessed rapid progress, particularly in the domain of text-to-image synthesis. Models such as Stable Diffusion are capable of producing photorealistic visuals from natural language prompts, enabling applications in digital art, content creation, and therapeutic tools. Among these, *visual therapy*—which involves using emotionally resonant images to support mental well-being—represents a promising frontier. However, current generative systems often lack personalization, limiting their emotional impact and therapeutic value.

To address this gap, we explore a novel approach to *personalized visual therapy* using DreamBooth, a fine-tuning technique that adapts pre-trained diffusion models to new concepts from a small number of images. Specifically, we fine-tune a Stable Diffusion model on custom datasets of personal pet images to generate context-aware visuals that are aligned with simple text prompts (e.g., "my dog sitting on a mountain"). By integrating CLIP-based natural language conditioning and latent space encoding, our pipeline generates emotionally relevant, high-fidelity images personalized to the user.

This paper makes the following contributions:

- We present a streamlined DreamBooth-based personalization pipeline for adapting Stable Diffusion to small-scale, user-specific image sets.
- We implement a lightweight, interactive deployment interface using Streamlit to enable real-time user interaction with the personalized generative model.
- We conduct both quantitative (CLIPScore and SSIM) and qualitative evaluations to assess semantic alignment and visual fidelity, identifying strengths and limitations in generation quality.

Our results demonstrate the potential of personalized diffusion models for empathetic applications like visual therapy, while also outlining current challenges such as dataset limitations and hallucination artifacts.

II. Background

Image synthesis has seen rapid progress with the development of diffusion models. These models generate images by progressively refining random noise through a learned reverse process. However, directly modeling high-dimensional image space can be computationally prohibitive. Techniques such as latent space compression—achieved through autoencoders like `AutoencoderKL`—address this challenge by encoding images into a lower-dimensional space. Moreover, recent advances in natural language processing have enabled conditioning generative models on text. By embedding text into a joint visual–language space (using models such as CLIP), generated images can be tailored to reflect specific semantic content.

The convergence of these techniques provides a powerful framework for personalized image synthesis, as explored in our work.

III. Related Work

This section reviews the foundational research that informs our approach to personalized image synthesis. We build upon recent advances in diffusion models, latent space compression, text-based conditioning, and image evaluation techniques. By combining these components, our work aims to enable emotionally resonant visual generation tailored to individual users.

A. Diffusion-Based Generative Modeling

Diffusion models have become a prominent framework for image synthesis. Ho *et al.* [1] introduced *Denoising Diffusion Probabilistic Models* (DDPMs), in which a forward process adds Gaussian noise to an image and a learned reverse process removes it. This iterative denoising strategy has shown strong performance in terms of image fidelity and training stability. Later improvements by Dhariwal and Nichol [2] further enhanced sample quality, making diffusion models a compelling alternative to GAN-based approaches.

B. Latent Space Compression via Autoencoding

Generating images directly in high-dimensional pixel space is computationally expensive and often inefficient. Latent Diffusion Models (LDMs) address this by encoding images into a lower-dimensional latent space using autoencoders such as `AutoencoderKL`. This strategy, as outlined by Rombach *et al.* [3], enables faster training and inference while preserving essential visual features. In our pipeline, pet images are first encoded into this latent space before undergoing the diffusion process, allowing the model to focus on meaningful content.

C. Natural Language Conditioning with CLIP

Radford *et al.* [4] introduced CLIP, a large-scale multimodal model trained on image–text pairs, which learns a shared embedding space for vision and language. This innovation enables generative models to condition on textual input by aligning image features with their semantic descriptions. In our work, we use CLIP embeddings to condition the diffusion process on text prompts, ensuring that generated outputs reflect both visual fidelity and semantic intent.

D. U-Net Architecture for Image Generation

The U-Net architecture, introduced by Ronneberger *et al.* [5], has become a core component in many image generation pipelines. Its encoder–decoder structure with skip connections preserves spatial details during upsampling, making it ideal for reconstruction tasks. In Stable Diffusion, the `UNet2DConditionModel` implements this architecture to iteratively remove noise from latent representations. We adopt this model as the denoising backbone in our system.

E. Reference-Free Evaluation with CLIPScore

Evaluating generated images without ground-truth references presents a unique challenge. Hessel *et al.* [6] proposed CLIPScore, a reference-free metric that measures semantic similarity between text prompts and generated images using CLIP embeddings. This allows for quantitative assessment of alignment between user intent and model output. Given the absence of paired image–caption data in our setting, CLIPScore serves as a practical and effective evaluation tool.

F. Discussion

The convergence of diffusion modeling, latent compression, language-guided generation, and semantic evaluation has enabled significant progress in personalized image synthesis. Our system integrates these elements into a unified pipeline for generating pet-specific images based on textual descriptions. By incorporating CLIPScore as a lightweight, reference-free evaluation method, we can assess output quality while maintaining system efficiency. This synthesis of methods supports the emerging potential of generative models in therapeutic and emotionally responsive applications.

IV. Methodology

Our proposed pipeline integrates a fine-tuned Stable Diffusion model with text-guided personalization for visual therapy. This section outlines our data preparation steps, training strategy, conditioning method, and evaluation metrics.

A. Data Collection and Preprocessing

We collected a small dataset of personal pet images from multiple sources. Each image was resized to 512×512 pixels, center-cropped, normalized, and tokenized using the `CLIPTokenizer` to enable text conditioning. These preprocessing steps ensure consistency across inputs and compatibility with the pre-trained model architecture.

B. Model Training

We fine-tuned a pre-trained Stable Diffusion model using the DreamBooth method, which allows adaptation to novel concepts with limited training data. The training pipeline includes:

- **Latent Encoding:** Images are encoded into a lower-dimensional latent space using `AutoencoderKL`, reducing computational overhead while preserving key features.
- **Denoising:** A `UNet2DConditionModel` is trained to reverse the noise addition process within the latent space. The U-Net architecture enables the model to maintain both global and fine-grained visual structure.
- **Diffusion Scheduling:** The noise addition and removal process is governed by a `DDPMScheduler`, which defines the forward and reverse steps of the diffusion process.
- **Optimization:** Training employs mixed precision, gradient checkpointing, and the 8-bit `AdamW` optimizer (via `bitsandbytes`) to optimize memory and performance.

We use the Hugging Face `Accelerate` framework for managing training workflows, supporting both multi-device distribution and efficient memory utilization.

C. Natural Language Conditioning

Text prompts are embedded using a CLIP-based encoder, mapping them into a joint visual–language latent space. These embeddings condition the diffusion model during generation, ensuring that the output aligns with the semantic content of the prompt (e.g., "my dog at the beach") while retaining the pet's visual identity.

D. Evaluation Metrics

We evaluate model performance using both quantitative and qualitative methods:

- **CLIPScore:** A reference-free metric that measures semantic alignment between generated images and their text prompts using cosine similarity in CLIP embedding space.
- **SSIM (Structural Similarity Index):** Assesses visual similarity between generated images and original inputs, focusing on luminance, contrast, and structural fidelity.

Together, these metrics offer a comprehensive view of both semantic coherence and perceptual quality.

V. Experiments and Results

In this section, we detail our experimental setup, present quantitative and qualitative evaluations, and discuss the end-to-end project flow—including both the overall process and the specific deployment flow via our Streamlit web interface.

A. Experimental Setup

Our experimental framework is implemented in Google Colab, leveraging GPU resources and mounting Google Drive to store the dataset and model checkpoints. We begin by installing the necessary packages (e.g., `diffusers`, `accelerate`, `bitsandbytes`, `triton`) and select an appropriate precompiled version of `xformers` based on the detected GPU type (T4, P100, V100, or A100).

The model training is based on a DreamBooth fine-tuning approach using a pre-trained Stable Diffusion model. Our custom dataset—comprising personal pet images—is preprocessed using standard transforms (resize, crop, normalization) and tokenized using a `CLIPTokenizer`. Key hyperparameters include:

- Learning Rate: 5×10^{-6}
- Maximum Training Steps: 300
- Batch Size: 2 (with gradient accumulation over 2 steps)
- Mixed Precision: FP16 with gradient checkpointing
- Optimizer: 8-bit AdamW (via `bitsandbytes`)

Training is conducted using Hugging Face's `Accelerator` to manage multi-device training, mixed precision, and efficient memory usage. During each training iteration, images are encoded into latent space using `AutoencoderKL`, noise is added via a diffusion scheduler (`DDPMScheduler`), and a U-Net–based model (`UNet2DConditionModel`) learns to reverse the noise addition.

B. Quantitative Evaluation

For evaluation, we rely on two metrics:

- **CLIPScore:** This reference-free metric computes the cosine similarity between image and text embeddings to quantify semantic alignment.
- **Structural Similarity Index (SSIM):** This metric is used to compare generated images against ground-truth images.

Table I presents the average CLIPScore and SSIM scores for various text prompts. Our experiments indicate that the generated images achieve high semantic consistency and visual fidelity.

TABLE I
Average Evaluation Metrics for Selected Prompts

Prompt Description	CLIPScore (%)	SSIM (%)
Pet at the Beach	63.92	28.66
Pet in a Superhero Costume	78.5	37.3

C. Qualitative Analysis

Qualitative assessments are performed by visually comparing generated images to highlight both the strengths and limitations of our approach. Figure 1 illustrates two examples:

- **Good Generation:** The image generated from the prompt "a photo of `<bruno>` in a park" closely resembles the images in the training dataset, demonstrating coherent details and semantic alignment.
- **Bad Generation:** The image generated from the prompt "a photo of `<bruno>` swimming underwater" exhibits logical inconsistencies and hallucinatory artifacts, indicating challenges in generalizing to out-of-distribution prompts.

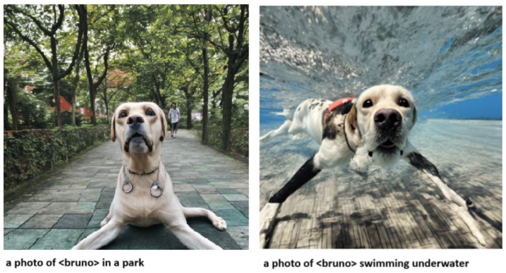

a photo of <bruno> in a park
a photo of <bruno> swimming underwater

Fig. 1. Combined qualitative example: The left portion represents a good generation for the prompt "a photo of `<bruno>` in a park" (resembling training images), while the right portion shows a bad generation for the prompt "a photo of `<bruno>` swimming underwater" with logical inconsistencies and hallucinations.

D. Project Flow Diagram

Figure 2 illustrates the overall project flow, capturing the main stages of:

1) **Data Preprocessing:** Personal pet images are collected, preprocessed, and tokenized.
2) **Model Training:** A pre-trained Stable Diffusion model is fine-tuned using the DreamBooth approach with latent space encoding, diffusion scheduling, and U-Net–based denoising.
3) **Inference and Evaluation:** The fine-tuned model generates images based on text prompts, which are then evaluated using CLIPScore and SSIM metrics.
4) **Deployment:** A user-friendly Streamlit web interface is deployed for real-time interaction.

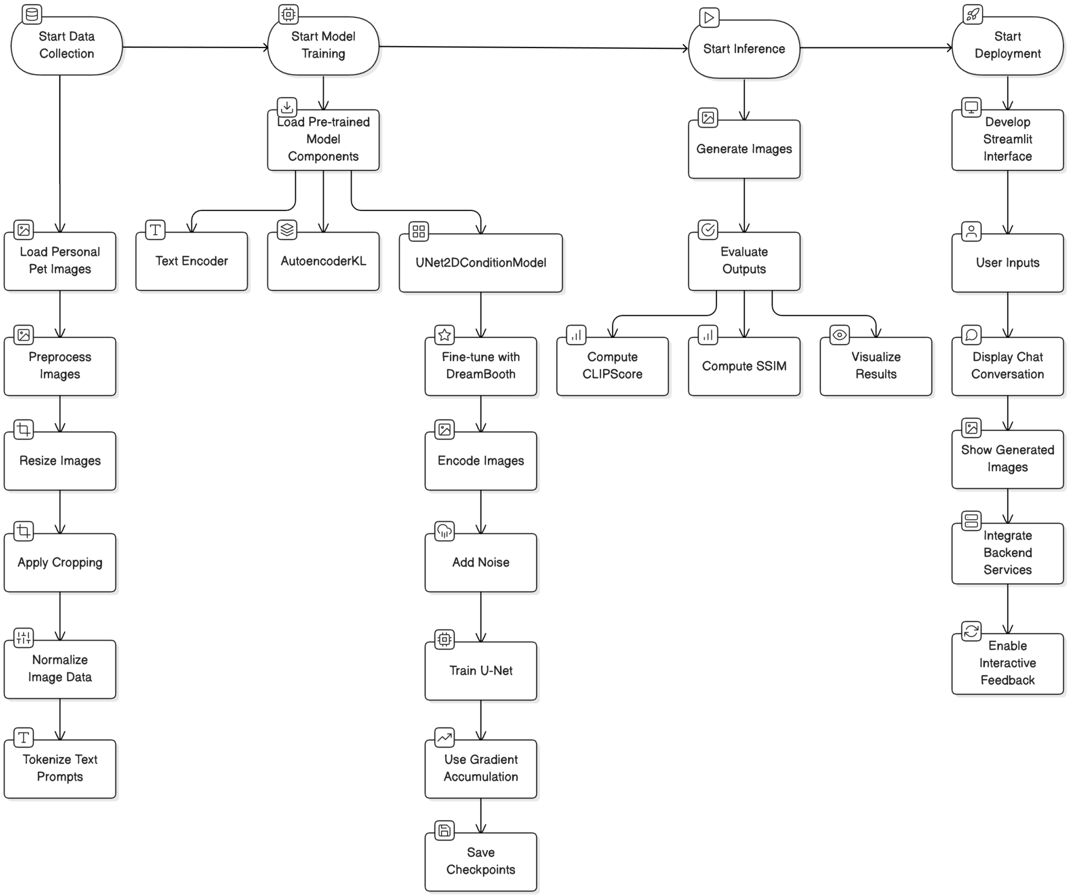


Fig. 2. Project Flow Diagram showing Data Preprocessing, Model Training, Inference/Evaluation, and Deployment.

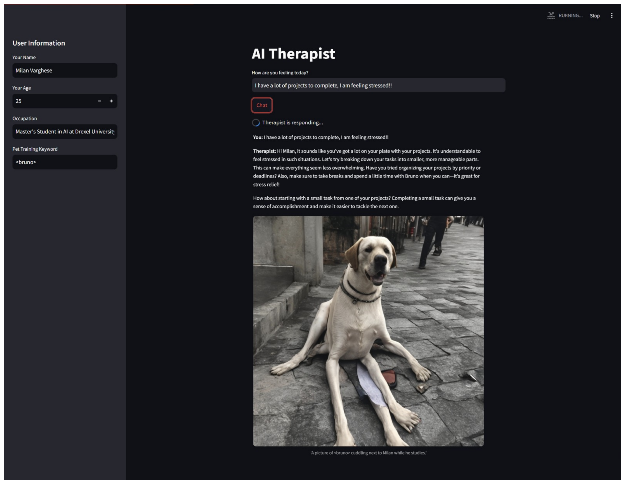
E. Deployment Flow Diagram

The deployment process involves:

1) **User Interface Access:** The user opens the Streamlit web application.
2) **User Input:** The user enters personal details (name, age, occupation, pet keyword) and describes their current mood.
3) **Submission:** The user clicks the “Chat” button, submitting their input.
4) **Backend Processing:** The input is processed by the Therapist Agent, which generates an empathetic text response and a concise image prompt.
5) **Image Generation:** The fine-tuned Stable Diffusion model generates an image based on the image prompt.
6) **Response Aggregation:** The Therapist Agent combines the text response and generated image.
7) **Display:** The Streamlit interface displays the combined output, updating the chat history.

F. Streamlit Website Interface and Output

To further demonstrate the system in action, Figure 3 shows a screenshot of the deployed Streamlit web interface. The interface displays the user input fields, chat conversation, and generated images. This visual output highlights the interactive nature of our deployment, where users can enter their details and mood, and receive an empathetic response along with a custom-generated image.



User Information

Your Name: Milan Varghese

Your Age: 25

Occupation: Master's Student in AI at Drexel Universit

Pet Training Keyword: <bruno>

AI Therapist

How are you feeling today?: I have a lot of projects to complete, I am feeling stressed!!

Therapist is responding...

You: I have a lot of projects to complete, I am feeling stressed!!

Therapist: Hi Milan, it sounds like you've got a lot on your plate with your projects. It's understandable to feel stressed in such situations. Let's try breaking down your tasks into smaller, more manageable parts. This can make everything seem less overwhelming. Have you tried organizing your projects by priority or deadlines? Also, make sure to take breaks and spend a little time with Bruno when you can—it's great for stress relief!

How about starting with a small task from one of your projects? Completing a small task can give you a sense of accomplishment and make it easier to tackle the next one.

'A picture of <bruno> cuddling next to Milan while he studies.'

Fig. 3. Screenshot of the Streamlit web interface showing user input and generated output.

VI. Project Limitations

While our system demonstrates promising results in personalized image synthesis, several limitations remain.

A major limitation is the relatively small size of the fine-tuning dataset, which restricts the model's ability to generalize. This can lead to overfitting and inconsistent generation quality, particularly for prompts that deviate from the training domain.

Additionally, variability in output quality remains a challenge. While some generations exhibit semantic alignment and strong coherence, others suffer from hallucinated features, distorted proportions, and poor object realism. These artifacts are more pronounced with abstract or imaginative prompts.

These limitations underscore the need for expanding the dataset, refining fine-tuning strategies, and incorporating stronger regularization techniques to enhance output stability and realism.

VII. Conclusion

This paper has presented a DreamBooth-based approach to fine-tuning a pre-trained Stable Diffusion model for personalized visual therapy. By leveraging a small, custom dataset of personal pet images, our system enables the generation of identity-preserving, context-aware visuals from simple text prompts. We integrated key techniques—including latent space compression, natural language conditioning with CLIP, and U-Net–based denoising—to construct a pipeline capable of generating semantically aligned and emotionally resonant images.

Our evaluation using both CLIPScore and SSIM metrics, alongside qualitative comparisons, confirms the effectiveness of our approach. The successful deployment of an interactive Streamlit interface further demonstrates the potential of personalized diffusion-based models in therapeutic applications.

VIII. Future Work/Extensions

Future research may explore several avenues for further improvement:

- **Enhanced Evaluation:** Incorporate additional metrics such as Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) to provide a more comprehensive assessment of image quality.
- **Dataset Expansion:** Augment the dataset with a larger and more diverse collection of images to improve the model's generalization and robustness.
- **Real-Time Implementation:** Optimize the training and inference pipelines to reduce latency and support real-time user interaction.
- **User Feedback Integration:** Introduce a feedback loop that allows users to rate or refine outputs, enabling iterative improvements and adaptive personalization.
- **Prompt Generalization:** Explore techniques such as prompt engineering or auxiliary fine-tuning to improve handling of abstract or out-of-distribution prompts.

Acknowledgment

The authors would like to thank Dr. Edward Kim for teaching CS 614: Applications of Machine Learning at Drexel University. The course's foundational concepts and material were instrumental in shaping the direction of this final project.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems, 34*, 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and D. Amodei, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[6] J. Hessel, A. Holtzman, M. Iyyer, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08736, 2021.

[7] Y. Song, C. Meng, and S. Ermon, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.

[8] "Streamlit documentation," https://docs.streamlit.io, 2020, accessed: 2025-03-21.

[9] "Accelerate: A deep learning training and inference library," https://github.com/huggingface/accelerate, 2021, accessed: 2025-03-21.