Speech to Image Generation Using Fine-tuned Latent Diffusion Model

Adil Qureshi
CECS, University of Michigan-Dearborn
Dearborn, United States
adilq@umich.edu

Shubham Jagtap
CECS, University of Michigan-Dearborn
Dearborn, United States
jshubham@umich.edu

Sudan Jaskirat
CECS, University of Michigan-Dearborn
Dearborn, United States
jsudan@umich.edu

***Abstract*—We propose a pipeline for fine-tuning Stable Diffusion v2 [2] using personalized image data and OpenAI Whisper for transcription-driven prompt generation. The pipeline utilizes DreamBooth [1] to fine-tune the model on a limited dataset of the subject, carefully curated to capture variations in pose, expression, and lighting conditions, along with class images representing the general category of the subject. Fine-tuning incorporates prior preservation loss to balance generalization and specialization of the model. The Whisper model is used to transcribe audio into textual prompts, enabling dynamic and contextually rich image generation. The fine-tuned model is trained using pretrained Stable Diffusion v2 with mixed-precision techniques for efficiency, and evaluated through qualitative and quantitative assessments. Results demonstrate that the model can generate highly consistent and realistic images of the subject across diverse scenarios and styles, showcasing the potential of combining personalized datasets with ASR systems for advanced generative AI applications.**

***Index Terms*—Stable Diffusion, speech-to-image generation, fine-tuning.**

I. Introduction

The rapid advancement of generative models, particularly in the field of text-to-image synthesis, has led to significant breakthroughs in the creation of highly realistic images from textual descriptions. Stable Diffusion is one such generative model that uses a diffusion-based architecture to produce high-quality images based on textual prompts. It operates by iteratively denoising random noise into an image, guided by the semantics of the given text. The model's ability to generate images that reflect complex and nuanced details, such as specific objects, environments, and artistic styles, has made it widely popular. Despite its impressive capabilities, a major challenge remains when attempting to generate highly personalized content with limited data, particularly when the subject's identity, pose, and context need to be preserved in a diverse range of generated images.

To address this, DreamBooth was introduced as a fine-tuning technique that allows a pre-trained generative model like Stable Diffusion to be adapted to new, specific subjects using a small set of images (often as few as 3-10). DreamBooth works by leveraging a method known as prior preservation loss, which ensures that while the model learns to generate images of the specific subject, it does not forget the generalization learned from its pre-training. This allows the model to generate highly personalized outputs while maintaining its ability to generate images of other objects, scenes, or people. During fine-tuning, the model learns detailed representations of the subject's features, such as facial structure, clothing, and typical environments, enabling it to generate consistent and accurate representations of the subject in new contexts and settings.

In this work, we build on the DreamBooth methodology by fine-tuning Stable Diffusion v2 with a carefully curated dataset consisting of only 15 images of a unique subject. The dataset is designed to include a variety of poses, expressions, and lighting conditions, ensuring the model learns the subject's distinct features without overfitting. Additionally, we enhance the image generation process by integrating OpenAI Whisper, an automatic speech recognition (ASR) system that transcribes audio prompts into text. This enables the generation of dynamic and contextually rich prompts from user-provided audio, facilitating a more interactive and versatile approach to image synthesis.

Our approach aims to combine personalized generative models with real-time, natural language-driven image generation. The fine-tuning process involves using prior preservation loss to maintain the balance between subject-specific features and general knowledge. Additionally, we optimize the training with mixed-precision techniques, which reduce memory usage and computational overhead, making the pipeline more efficient. Through qualitative and quantitative evaluations, including CLIP-based similarity metrics, we assess the effectiveness of this pipeline in generating consistent, high-fidelity images that accurately represent the subject across diverse scenarios. By leveraging DreamBooth with Stable Diffusion and Whisper, this work demonstrates the potential of personalized generative models in creating realistic, creative, and interactive content.

II. LITERATURE SURVEY

A. Latent Diffusion Model

Stable Diffusion v2 is a cutting-edge text-to-image generative model based on the concept of latent diffusion, introduced by Rombach et al. [1]. It operates in a lower-dimensional latent space, allowing for more computationally efficient image generation compared to traditional pixel-based methods. The

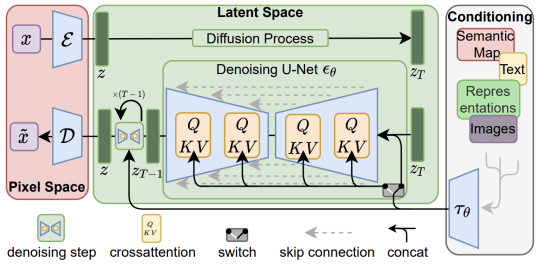

Fig. 1. Stable Diffusion Architecture.

model is trained using a denoising score-matching objective, where the task is to iteratively reverse a noising process applied to an image, conditioned on textual prompts. The architecture is shown in Figure 1.

At its core, Stable Diffusion is built on a U-Net architecture, a type of neural network commonly used for image segmentation tasks, but in this context, adapted for image generation. The architecture comprises an encoder-decoder structure with skip connections between corresponding layers of the encoder and decoder. These skip connections are particularly important because they allow the model to retain detailed spatial information as it processes the input data. The U-Net in Stable Diffusion is used during the denoising process in the reverse diffusion steps. The U-Net in Stable Diffusion is central to its ability to process high-dimensional data in an efficient manner. The network consists of two main parts: an encoder that progressively downscales the input and a decoder that progressively upscales the latent representation back to the original image resolution.

Stable Diffusion's efficiency comes from the use of latent spaces, where images are compressed into a more compact representation before undergoing the diffusion process. This reduces computational complexity and memory usage, allowing the model to generate high-quality images without the heavy computational load associated with pixel-level diffusion models. The model has shown impressive performance across diverse datasets and prompts, demonstrating its ability to handle complex image generation tasks with minimal computational resources.

B. DreamBooth

DreamBooth is a fine-tuning technique introduced by Chou et al. [1] that adapts pre-trained generative models like Stable Diffusion to specific subjects using a minimal number of training images, typically 15 images. DreamBooth allows the model to specialize in generating images of a unique subject while preserving its ability to generalize to other subjects and objects. This is achieved through the introduction of a prior preservation loss, which ensures that the model maintains its generalization ability while focusing on learning the specific features of the subject.

The DreamBooth fine-tuning process involves two main components and as shown in Figure 2 :

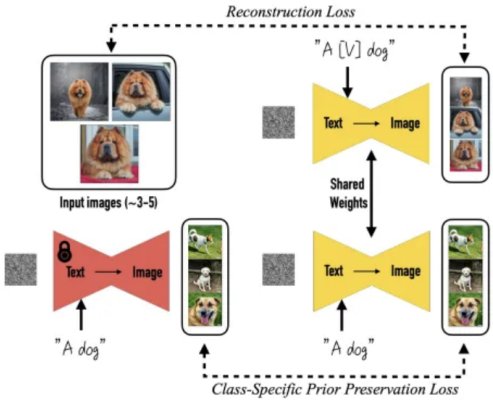

Fig. 2. Fine-tuning process: Starting with 3–20 images of a subject, we fine-tune a text-to-image diffusion model. The input images are paired with a text prompt that includes a unique identifier and the class name to which the subject belongs.

- Instance-specific adaptation, where the model learns the unique features of the subject (e.g., facial features, clothing, or specific pose) from the small dataset.
- Class-preserving regularization**, which allows the model to continue generating diverse outputs for more general categories, such as "person" or "dog," by using class images or class-conditioned prompts.

This technique enables the model to retain its general knowledge and flexibility while generating highly personalized content. By leveraging this method, DreamBooth achieves efficient adaptation to new concepts without overfitting to a small dataset, making it suitable for applications where collecting large amounts of training data is impractical.

C. OpenAI Whisper

Whisper, developed by OpenAI [4], is a robust automatic speech recognition (ASR) model capable of transcribing spoken language into text. Whisper is trained on a massive multilingual and diverse audio dataset, which allows it to handle a wide range of accents, background noise, and varying audio conditions. The model uses a transformer-based architecture, which has been highly effective in natural language processing tasks, and it operates by first encoding the audio input into feature sequences, which are then processed by the transformer model to generate corresponding text.

One of the key advantages of Whisper over traditional ASR systems is its versatility. It can transcribe speech in a wide array of languages and domains, making it a useful tool for a variety of applications. Whisper's ability to perform in challenging real-world scenarios—such as noisy environments and diverse accents—sets it apart from earlier ASR systems that required domain-specific fine-tuning.

In the context of generative models like Stable Diffusion, Whisper offers a powerful integration by enabling dynamic text generation from audio. This allows users to interact with the model by simply speaking, rather than typing out prompts. By transcribing audio into text, Whisper enables a

more natural, real-time method for generating images based on spoken descriptions. This integration bridges the gap between speech and visual generation, offering a more intuitive user experience.

III. Methodology

A. Dataset

For this project, Stable Diffusion v2 is pre-trained on the LAION-5B dataset, which includes around 5 billion image-text pairs. This dataset spans a wide range of domains, including landscapes, objects, animals, and people, enabling the model to learn the relationships between text and images. The generalization achieved from this diverse training set allows Stable Diffusion v2 to generate high-quality images from textual descriptions.

To fine-tune the pre-trained model for personalized image generation, we collect a small dataset of 15 images of a specific subject (e.g., a person in our case) that represent the subject from different perspectives, lighting conditions, and poses. These images are curated to cover a wide range of variations of the subject, ensuring the model is exposed to the subject in diverse scenarios. The images are then preprocessed to ensure uniform resolution and quality, with all images resized to a consistent resolution of 512x512 pixels.

B. Flow chart

Figure 3 illustrates the system flowchart, outlining the overall workflow of the proposed method. The process is divided into two key phases: Phase 1 involves converting speech to text, while Phase 2 focuses on generating images from the processed text. The flowchart details each step, showcasing how the system effectively produces personalized and context-specific outputs.

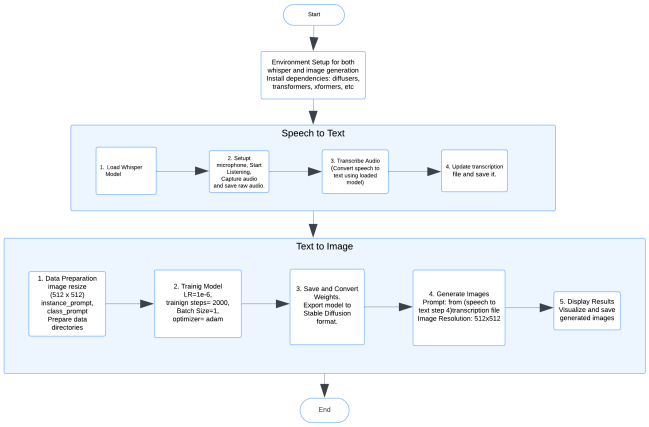

Fig. 3. System flow chart.

C. Fine-Tuning Process

The fine-tuning process involves adapting the model to generate personalized images of the subject in our dataset by freezing the rest of layers other than last few layers. We use the DreamBooth framework to perform the fine-tuning, which allows us to introduce a new concept (the subject) into the model's latent space.

The training involves setting hyperparameters such as a learning rate of `1e-6`, a batch size of `1`, and a gradient accumulation of `1`. The training is performed with mixed precision (`fp16`) and uses the Adam optimizer for memory-efficient optimization. The learning rate is kept constant throughout the training process using a constant learning rate scheduler. The model is trained for a maximum of 2000 steps, with images being saved at every 500-step interval for visual evaluation. The training process is carried out on a single GPU, with each training step generating a sample image based on a prompt such as "photo of jaskirat person in forest".

D. Evaluation

After fine-tuning, the model is evaluated by generating images based on textual prompts describing the subject. The evaluation focuses on the model's ability to generate accurate and contextually appropriate images of the subject from various prompts. Generated images are compared visually to assess how well the model has learned to replicate the subject in different contexts. The evaluation is performed both qualitatively (by human assessment of the generated images) and quantitatively (by monitoring loss) Figure 4.

In addition to generating images of the subject, we also assess the model's ability to maintain diversity and avoid overfitting to the specific characteristics of the training images. The generated images should exhibit variations in pose, lighting, and background, reflecting the generalization capabilities of the fine-tuned model.

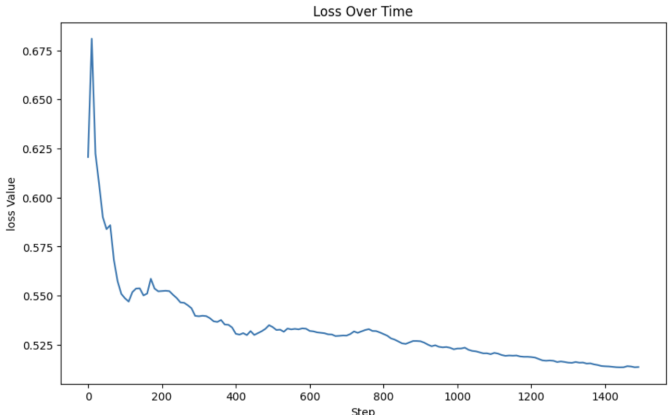

Fig. 4. Model Loss

IV. Results

The following images in Figure 5. demonstrate the model's ability to generate contextually accurate and visually coherent outputs based on prompts. The model successfully interprets diverse prompts involving complex activities, environments, and individual identities. Specifically, the results showcase the model's ability to synthesize realistic images of individuals in

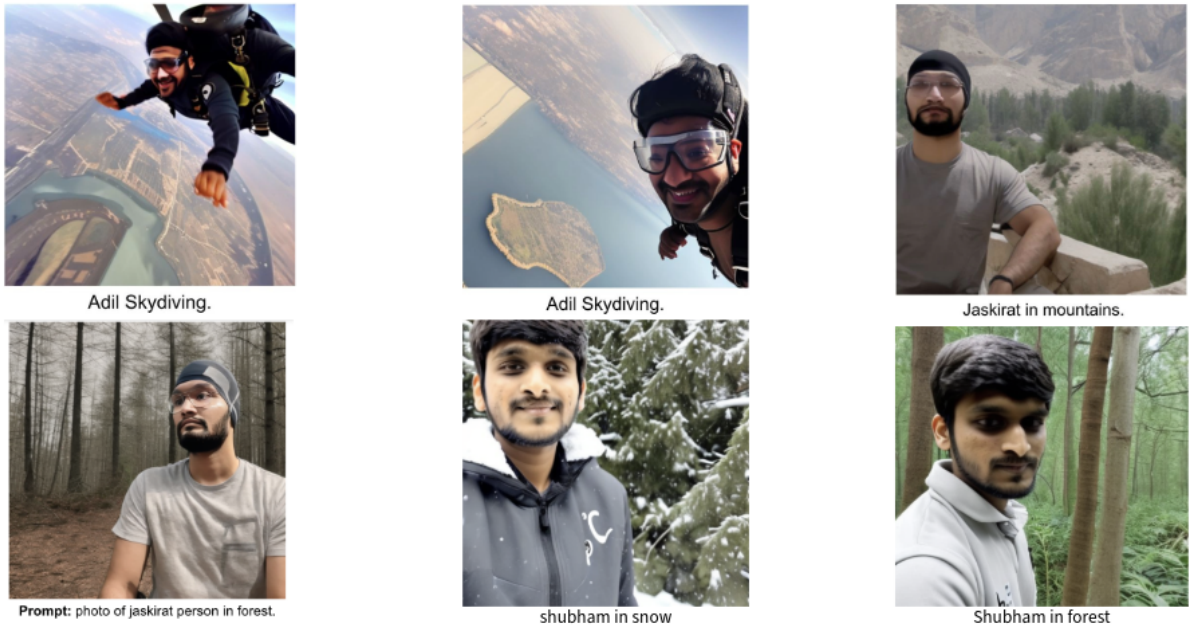

Fig. 5. Images generated by model

various settings, such as extreme activities like skydiving and natural landscapes like mountains or forests. This indicates the model's potential in generating customized visual content for applications such as creative media, content generation, and other domains requiring context-aware imagery.

REFERENCES

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[2] A. Blattmann, R. Rombach, K. Oktay, and B. Ommer, "Retrieval-Augmented Diffusion Models," arXiv, 2022, doi: 10.48550/ARXIV.2204.11824.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10674-10685, doi: 10.1109/CVPR52688.2022.01042.

[4] OpenAI. (2022). Whisper: Robust speech recognition across diverse languages. Available at: https://openai.com

[5] Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. Hoi, "BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding and Generation," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[7] A. Ruiz, T. Li, S. Ebrahimi, E. Carlson, and R. Salakhutdinov, "DreamBooth: Fine-tuning Text-to-Image Diffusion Models for Subject-Driven Generation," arXiv preprint arXiv:2208.12242, 2022.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, and others, "Photorealistic Text-to-Image Diffusion Models with Imagen," arXiv preprint arXiv:2205.11487, 2022.