# GhibliDream: Fine-Tuning Diffusion Models in Ghibli Style

Members: Aditya Potnis, Aryaman Nasare, Kartik Ramesh, Ridhwik Shravan Kalgaonkar
NetIDs: apotnis2, nasare2, kartikr2, rk44

Text-to-image diffusion models like Stable Diffusion demonstrate impressive generalization across diverse visual concepts. However, they often struggle to consistently replicate niche artistic styles such as the hand-painted aesthetic of Studio Ghibli. Training such models from scratch is computationally infeasible. In this project, we investigate lightweight fine-tuning approaches, specifically DreamBooth, LoRA, and full-parameter finetuning—to adapt pre-trained diffusion models using small, curated datasets.

After comparing the trade-offs across these methods, we focus on DreamBooth due to its effectiveness in few-shot adaptation and low compute requirements. We curated Ghibli-style image-caption pairs into task-specific subsets, manually enhancing captions for stylistic alignment. Our evaluations combine qualitative analysis and embedding-based similarity metrics (CLIP, DINOv2) to assess the visual fidelity and semantic consistency of outputs.

We observe that DreamBooth enables robust stylization while preserving base model capabilities. Experiments with unique token identifiers, sampling configurations, and multi-stage fine-tuning show that style consistency and subject-background blending can be significantly improved through careful prompt engineering and dataset composition.

Images generated using fine-tuned model

# 1. Introduction

Modern text-to-image diffusion models like Stable Diffusion have shown remarkable capabilities in generating visually coherent and semantically aligned images from textual prompts. However, these models often falter when asked to reproduce highly specific or niche artistic styles, such as the iconic hand-painted aesthetic seen in Studio Ghibli films. This inconsistency arises from the broad and generalized nature of the training data, which may not capture the subtleties of such visual styles.

Achieving stylistic consistency is important for use cases in animation, gaming, personalized art generation, and AI-assisted creative workflows. However, fine-tuning diffusion models for this purpose presents practical challenges. Full retraining from scratch is prohibitively expensive due to the large number of parameters and high GPU memory requirements. Moreover, acquiring large, high-quality, labeled datasets in niche styles is often infeasible.

Prior research has explored various methods for adapting diffusion models to specific subjects or styles without retraining from scratch. Notable among these are:

1. **Full-parameter fine-tuning**, which updates all model weights but demands significant compute.
2. **LoRA (Low-Rank Adaptation)**, which injects small trainable adapter layers to reduce training cost.
3. **DreamBooth**, a technique that associates a unique token with a new visual concept or style, enabling few-shot fine-tuning with prior preservation.

In this project, we investigate how to adapt Stable Diffusion models to generate images in a consistent Ghibli style using small, curated datasets. After experimenting with all three methods, we found DreamBooth to offer the best trade-off between efficiency, quality, and accessibility. Our approach involves:

1. Curating subsets of Ghibli-style images and enhancing their captions for better prompt alignment.
2. Fine-tuning Stable Diffusion with DreamBooth using unique token identifiers and prior preservation.
3. Evaluating results through both qualitative inspection and quantitative similarity metrics (CLIP, DINOv2).

Our goal is to analyze how lightweight fine-tuning techniques can be leveraged to inject strong stylistic priors into pre-trained diffusion models, enabling high-quality, stylized image synthesis under real-world resource constraints.

# 2. Details of Approach

At the beginning of the project, we explored three fine-tuning methods in parallel: full-parameter fine-tuning, LoRA adapters, and DreamBooth. Our goal was to adapt Stable Diffusion models to specific artistic styles While each approach showed some potential, we consistently encountered resource limitations, particularly around GPU memory and storage capacity, which restricted our ability to train large or complex models.

All of our training was carried out through Colab (A100 and T4 instances) or through our local GPUs. Full-parameter fine-tuning gave us the most control over the model but quickly proved infeasible due to its high computational demands. It required substantial GPU memory and generated large checkpoints, making it unsuitable for our Colab-based setup and limited local hardware. LoRA adapters offered a more lightweight alternative by updating only a subset of model parameters. However, while LoRA reduced memory requirements, it still involved long training times and often produced inconsistent outputs, particularly when working with small or noisy datasets.

After running experiments across all three methods, we found that DreamBooth offered the best balance between quality and efficiency. It required significantly fewer images, trained faster, and produced more consistent results without the heavy computational overhead of the other approaches.  As a result, we chose to standardize our workflow around DreamBooth and focus on building small, well-curated datasets tailored to each style, which aligned well with both the method's strengths and our resource constraints.

## 2.1 Dataset Creation

We used the Nechintosh/ghibli dataset, available on HuggingFace, which contains 810 images sourced from Studio Ghibli's official gallery at ghibli.jp. Each image is paired with a caption generated using the BLIP2 model. The dataset is released for non-commercial use under terms specified by the source website. This dataset was originally used in Approach 2 of our experiments, as described in our progress update report. However, we observed that many of the auto-generated captions were vague or inaccurate, which limited their usefulness for prompt-based image generation and fine-tuning. To address this, we curated three smaller, task-specific subsets of the dataset with manually improved captions tailored to our style transfer goals.

Each curated subset included updated image captions with more descriptive and context-aware language. Captions were rewritten using either GPT-4o or Gemini, and a unique identifier was prepended to each caption to enable style tagging during training. The images were manually chosen from the ghibli dataset.

The first subset contained 10 randomly selected images with diverse content, including characters in action and background compositions. Captions were generated using GPT-4o and designed to follow the descriptive style of the John Singer Sargent (WikiArt) dataset. These

captions included detailed references to subjects, composition, color palette, and emotional tone of the artwork.

The second subset focused exclusively on scenery and background art. It included approximately 15 images with captions generated using Gemini. These captions emphasized visual detail, environment description, and atmospheric cues without referencing characters or narrative elements.

The third subset targeted facial imagery and included 27 images featuring clear and prominent faces. Captions were again generated using Gemini and aimed to enhance facial description, expressions, and contextual background. For our final output, we trained with around 17 images as it provided the best results.

To streamline this process, we developed a short script to automate prompt rewriting by using the Gemini API. All of the curated subsets used in our experiments have been included with our submission for reference.

Our Diffusion training dataset and models (needs illinois google account to access): Datasets

## 2.2 Training Process

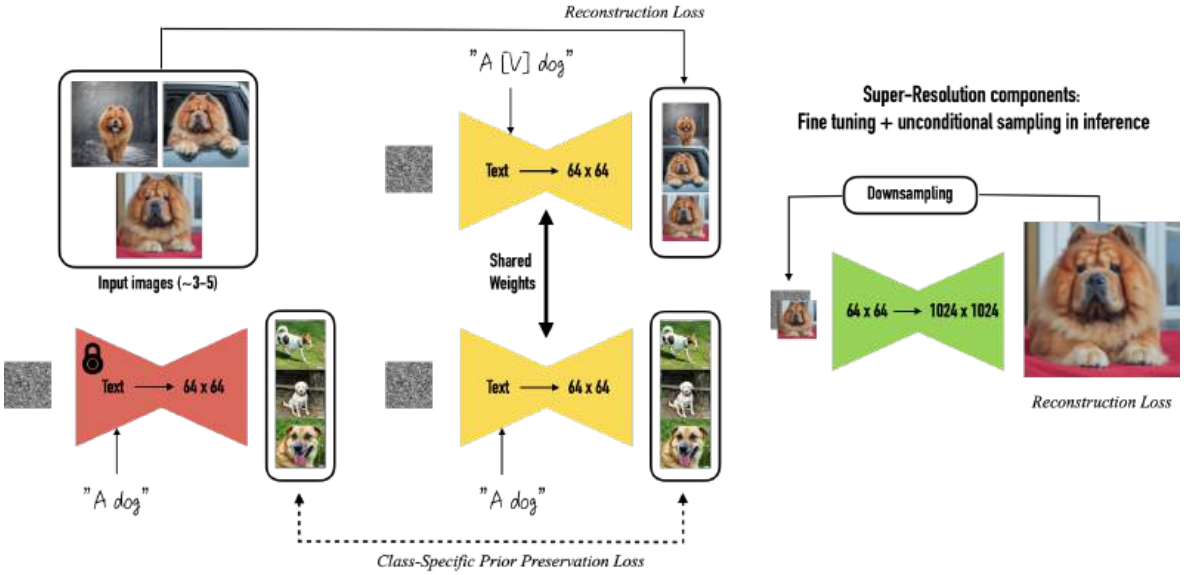
We utilize the Stable Diffusion 1.5 and 2.0 models to train on a smaller hand selected subset to test the performance of these systems with finetuning. In our early stages, we tested out a few combinations of Low Rank Adaptation, full training and other approaches for our progress update. In our initial tests, the LoRA fune tuning showed a lot of interesting trends, but we found that the dataset we required for good performance was quite large (200 - 400 images -caption pairs needed) and took a few hours for training for just 1500 epochs. (Tutorial: https://stabilityai.notion.site/Stable-Diffusion-3-Medium-Fine-tuning-Tutorial-17f90df74bce4c62a295849f0dc8fb7e)

| | | |
|---|---|---|
| LoRA finetuning experiment<br>Prompt:<br>Two characters standing in a garden in ghibli art style. |  <br>not fine-tuned |  <br>fine-tuned |

| Prompt | Sample Image |
| --- | --- |
| A woman in a blue dress with cats jumping around her, with a green locker in the background. |  |

Compared to LoRA, we also tested DreamBooth [1].

This approach works by associating a specific unique text token with the training / finetuning. The prompt can be like "sks person" or "xyz style". To prevent “forgetting” training also includes training for generic prompts to ensure that old relations are not lost in fine tuning.
This allows the system to fine tune with high accuracy for a few-shot system of size 3-5 images. In our experimentation, we found the model being able to finetune with a subset of the ghibli dataset with 10 to 17 image-text pairs. Dreambooth requires a unique identifier e.g. “a [V] dog” as used in the diagram. For our testing we experimented with a variety of training unique identifier tokens including training by staggering them for mixing of different features such as background and foreground merging from different movies or images. Dreambooth works by fine-tuning for a specific unique input while also keeping a prior preservation loss to prevent loss of the previous training weights in the process.




The training process for Dream Booth is a faster process overall compared to LoRA.
This can be attributed to the fact that unlike LoRA, Dreambooth uses a specialized prompt for fine tuning the Stable diffusion Layers. LoRA on the other hand injects additional layers between

the present layers, leading to training requiring more images in the dataset and longer epoch for fine tuning.

| | | | |
|---|---|---|---|
| SDXL 1.5 (Original) |  |  |  |
| DreamBooth (Fine-tuned) |  |  |  |
| Prompt | Cartoon of Scott Pilgrim and his friend Neil standing | Portrait of scott pilgrim | Cartoon of Jenny Chau from scott pilgrim movie |

While testing with DreamBooth, we discovered that the hair color for some characters on the smaller dataset trained on 300 epochs had blending of hair color of 2 different characters. Our hypothesis for this problem was that the model needed a longer training process and better captions to discriminate prompt features better. We decided to investigate this further after the progress update.

We decided that due to the lower compute requirements and faster training process, DreamBooth would be the focus on testing for the final part. We tested a variety of manually curated datasets which had foreground characters and different background images in the ghibli art style from different studio ghibli movies to verify the model. Additionally we also performed tests to verify staggered fine tuning with and compared with an ablation to check if fine tuning with multiple datasets can affect the stylization with two independent unique identifiers.
We used the synthetically generated dataset from section 2 to then further test our finetuning.

Further, we also analyzed the impact of sampling count, sampler model and image resolution settings and their effect on image outputs.

In the process of tuning we also found a quirk in the SD2 UNet sampling process using Gradio. As our system is trained to output a 512x512 image, when we tested this image in Gradio, we found that if we double the sample resolution to 1024x1024, the sd2 model experienced a tiling effect. The model attempted to use a 512 convolution for a 1024 image which led to a grid forming, but due to the probabilistic nature of the diffusion model, the grid was not identical but highly similar to the neighboring section, with the UNet feature upsampling in the model trying to compensate for the deviation and creating a unique set of esoteric and surreal art.

| | |
|---|---|
|  A girl with red hair in front of a leaf |  A house in the forest |

Tiling effect with similar but not same image quadrants, it compensates for the neighbor section and reduces noise.

# 3 Results

Video of the sampling process for images: Screen Recording 2025-05-13 134858.mp4

## 3.1 Qualitative results

We tested the model with a combination of caption lengths and found longer captions were better at encoding the scene than shorter captions. In our initial tests, images had poorer eye reconstruction. We found that as we scaled the dataset size, our training epochs had to be scaled as well. For most models since our samples were in the range of 10 - 20, the minimum epochs we needed were around 3000 and saw good results with reliable hand reconstruction at epoch 5000. We also found that the teeth and other features also improved with epochs.

Right image has a modified prompt with the hair color being requested as blonde and the color of dress to be changed to blue.

Prompt: A stylized anime illustration depicting a young girl walking alongside a vast, vividly detailed green leaf that dwarfs her figure, emphasizing a miniature or shrunken perspective. The girl, dressed in a deep red outfit with a cream-colored sleeve and a khaki strap crossing her chest, wears her auburn hair in a practical ponytail, fastened with what appears to be an oversized clothespin-like accessory, hinting at a whimsical or fantastical world where ordinary objects are reimagined as tools. Her downward gaze and solemn expression convey a sense of quiet determination or introspection. The lush, veiny leaf in the background is rendered with painterly strokes and subtle tonal variations of green, punctuated by occasional brown blemishes and illuminated by soft, diffuse light. The girl's shadow, cast crisply against the leaf, mirrors her posture and adds a grounding sense of realism. The overall composition evokes a serene yet adventurous atmosphere, blending natural wonder with a touch of magic.

## 3.2 Experiment methodology, CLIP and DINOv2 similarity metric

We used the implementation of FastDreamBooth by TheLastBen for our training process.
https://colab.research.google.com/github/TheLastBen/fast-stable-diffusion/blob/main/fast-DreamBooth.ipynb
Github: https://github.com/TheLastBen/fast-stable-diffusion?tab=readme-ov-file

Additionally we implement the dataset caption generation, similarity score measurement and other miscellaneous data handling scripts are shared on our repo and in the zip file.

For all the trained models, we used CLIP [3] image ViT embeddings as a metric for the image similarity score as CLIP has a 75.2% top1 score on imagenet. We used CLIP to analyze the embeddings to understand the latent space similarity to help quantify the semantic nature of the image in some way. We also utilized DINOv2 [2] to generate the embeddings with better spatial context for the global image context using the [CLS] token embedding. We then used these to perform L2 (Euclidean), Cosine and Manhattan distance metric similarity measurements between the sample training images with the generated images on similar prompts to quantify the model performance along with manual viewer verification.

We used different metrics to measure similarity scores, the variables x, y are the embedding vectors of input and reference values respectively in $R^n$ .

1) L2 / Euclidean distance:

$$d_{L2}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2) Cosine similarity:

$$s_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \, \sqrt{\sum_{i=1}^{n} y_i^2}}$$

3) Manhattan distance:

$$d_{L1}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

Listed below are the combination of the training methods and combinations of caption, image dataset type and unique identifier we used encoding.
Model training info:

| Model | Base model | epochs | Samples for training | Dataset name | LLM caption | Description (default: gh1b11 unique identifier used for text) |
|---|---|---|---|---|---|---|
| test-long-captions | SD2 512x512 | 5000 | 27 | ghibli-faces-l… | Gemini 2.0 flash | Trained on long text captions from movie scene images with faces of characters |
| test-long-captions-v2 | SD2 512x512 | 5000 | 17 | ghibli-faces-l… Removed 10 through 19 to reduce to 17 | Gemini 2.0 flash | Trained on long text captions from movie scene images with faces of characters |
| test-short-captions-v1 | SD2 512x512 | 5000 | 17 | ghibli-backgr… Removed 10 through 19 to reduce to 17 | Gemini 2.0 flash | Trained on short text captions from movie scene images with faces of characters |
| test3 | SD2 512x512 | Checkpoints 500, 1500, 2500, 3500, 4500 | 10 | general ghibli dataset<br><br>10 images 3 backgrounds, 7 characters | Openai gpt 4o | |
| test2 | SD1.5 512x512 | 3000 | 10 | general ghibli dataset<br><br>10 images 3 backgrounds, 7 characters | Openai gpt 4o | |
| ghibli-background-long | SD2 512x512 | 3000 | 15 | ghibli-backgr… | Gemini 2.0 flash | Trained on long text captions from movie scene images with the background. Used for 2nd stage fine tuning with faces |
| ghibli-background-short | SD2 512x512 | 3000 | 15 | ghibli-backgr… | Gemini 2.0 flash | Trained on short text captions from movie scene images with the background. |
| ghibli-faces-on-background | SD2 512x512 | 3000 | 5 | faces long dataset 5 | Gemini 2.0 flash | Trained on long text captions from movie scene images with faces of characters. Trained after training on the background |
| ghibli-faces-on-background-9hi6li | SD2 512x512 | 3000 | 5 | faces longdataset 5 9hi6li uniquetoken | Gemini 2.0 flash | Trained on long text captions from movie scene images with faces of characters. Trained after training on the background. ( 9hi6li unique identifier used for text prompt) |

## 3.3 Merge of background and character using successive finetune training

We merged 2 different types of fine tuning in the model to analyze output on how the unique DreamBooth identifier token affects model performance.
We performed 2 tests where we initially fine tuned a Stable Diffusion 2 model on 512x512 with 15 image samples as background images for 3000 epochs. The model generally is much better at background finetuning than characters, and we saw good image outputs. We took this fine tuned model and added a smaller sample of 5 images of ghibli characters and corresponding prompts and trained further for 3000 epochs.

We then performed two tests with the training data. We tested with identical identifiers for the background as well as non-identical identifiers in the text prompt section.

i) Identical (gh1b11 identifier common for background and foreground training)
ii) Non-identical (gh1b11 identifier for background and 9hi6li identifier foreground train)

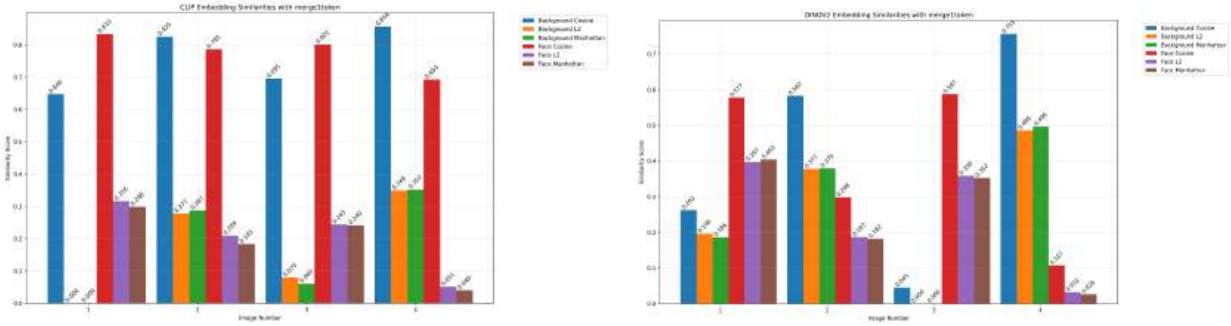
We found on visual inspection that the non identical identifier tokens model performed better at adding the subject in the requested background. The overall style transfer was also consistent between 2 characters which may come from different movies. For the identical movie dataset, the quality was generally a bit on the poorer side as the model seemed to focus more on the background than the subject.
We theorize that the foreground subject was described with fewer words than the background combined with the larger sample size for background images, which led to the images having poorer quality in terms of subject placement and appearance. We also noticed that the diffusion process, due to its probabilistic nature, led to 20% of attempts of image generation with no subject and just a background image. We also found that with sufficient training, the system was able to better recreate hands provided the text description was sufficiently thorough.

We found that the CLIP cosine similarity scores for the face in the images were much closer to the original data in the two unique token setup. Whereas the background had higher similarity scores for the non unique token training setup. For DINOv2 embeddings, we saw that the embeddings' cosine similarity for the background had diverged for the two tokens method. This may be due to the DINO embedding focusing more on the spatial conformation of the model, which we saw was better for the background in the one shared unique token method, at the cost of the character generation quality.

| Test1 of identical tokens (Steps: 150, Sampler: Euler a, CFG scale: 7, Seed: 1569673953, Size: 512x512, Model hash: 3b501be892, Model: ghibli-faces-on-background) | | | |
|---|---|---|---|
| Input prompt (character desc) | Input prompt (background description)<br><br>Prompts are concatenated | Input sample image (character)<br>Input sample image | CLIP - image / DINOv2 cosine similarity |

| | | (background) Output image | score |
|---|---|---|---|
| gh1b11, The image is a medium shot of a young man, presumably a teenager, standing in front of a shelf of books. He has short, dark brown hair and dark brown eyes. He is wearing a white collared shirt. He has a slight smile on his face. | gh1b11, The image depicts a cluster of buildings nestled amongst lush greenery, possibly on a hillside or elevated area. A prominent structure is a two-story building, painted a pale yellow with green accents, situated near the top of the composition; atop it flies a flagpole with two flags, one white and red and one blue and white. Several other buildings, mostly with dark-colored roofing and lighter-colored walls, are interspersed among the trees and foliage. The overall scene is bathed in a warm, golden light, suggesting either sunrise or sunset. | Inputs   Output  | face CLIP: **0.853** face DINO: **0.577** background CLIP: **0.646** background DINO: **0.262** |
| gh1b11, The image shows a character with dark green, almost black, straight-cut bob with bangs, standing with pink and purple flowers. The character has fair skin, green eyes, and a gentle smile. They are wearing a white garment with blue trim. | gh1b11, The image is a painted scene of a lush garden leading to a gazebo with an ocean view. A stone wall covered in pink climbing roses is seen on the left, next to a six-sided wooden gazebo with a red-tiled roof and a round attic window. A person wearing a hat can be seen inside the gazebo looking out. A stone path with steps leads from the foreground towards the gazebo and the ocean. The garden is filled with various colorful plants, including reds, whites, purples, and yellows. Overhanging the garden from above are leafy green branches. | Inputs   output  | face CLIP: **0.785** face DINO: **0.298** background CLIP: **0.825** background DINO: **0.582** |
| gh1b11, The image shows an animated young woman standing in what appears to be a rainy environment. She is partially shielded by a large, broad green leaf, possibly from a plant like taro, which she is holding up by its stalk. | gh1b11, The image is a hand-painted style illustration of a three-story building, presumably an older structure with weathered brown wooden paneling. Each story features multiple windows with blue panes, some with framed decorative molding. A balcony with a stained-glass arched window is central to the second story. The balcony appears to have a weathered grey-painted trim | Inputs  | face CLIP: **0.801** face DINO: **0.587** background CLIP: **0.695** background DINO: **0.045** |

| | | | |
|---|---|---|---|
| The leaf has numerous clear water droplets on its surface. The woman has shoulder-length brown hair, with a wave on one side, fair skin, and light-colored eyes. She is wearing a top that is brown around the chest and shoulders, transitioning to a light yellow for the arms and lower portion. | and is supported by two grey pillars. The building is festooned with numerous vertical banners, some hanging crookedly. The banners are white with black Japanese text. Some appear to be protest signs. Laundry is hanging from the windows on the left side. There is greenery at the building's base, indicating it may be a ground-level view. The sky is lightly blue in the upper right corner, with ver |  <br>Output<br>  | |
| gh1b11, The image is a medium shot of a young character with dark, choppy hair, light brown skin, and wide, light-colored eyes. They have white markings on their cheeks. The character is wearing a brown, fur-like garment over a darker top, secured with a light yellow clasp at the neck. Their arms are covered with green bracers. They are holding a bow in one hand and raising the other hand, gesturing with three fingers extended. | gh1b11, The image depicts a serene and picturesque scene centered around a traditional Japanese house. The house has a dark, tiled roof with defined rows of tiles and a wooden frame that supports the roof's overhang. The house features shoji screens that obscure the view inside. A raised wooden platform serves as a porch, with a stone step leading up to it. The interior has a picture frame and shelves on the back wall. Outside, a garden surrounds the house, filled with round, lush green bushes and trees. A large pine tree, with its characteristic sprawling branches, looms behind the house, partially covered by a cherry tree in full bloom, with delicate white and light pink flowers. The sky in the background is a soft gradient of pink and blue hues, suggesting either dawn or dusk. The overall atmosphere is tranquil and peaceful. | Inputs<br>  <br>  <br>Output<br>  | face CLIP: **0.693**<br>face DINO: **0.107**<br>background CLIP: **0.856**<br>background DINO: **0.755** |

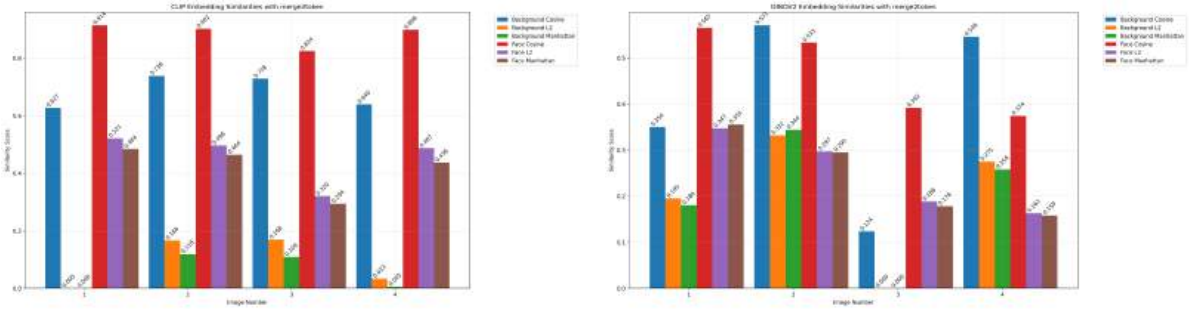


CLIP and DINOv2 embeddings for original input face and background images with respect to output for L2, Cosine and Manhattan similarity metrics. Tested on a shared unique identifier token.

| Test 2 of unique tokens (Steps: 150, Sampler: Euler a, CFG scale: 7, Seed: 1569673953, Size: 512x512, Model hash: 3b501be892, Model: ghibli-faces-on-background-9hi6li) | | | |
|---|---|---|---|
| Input prompt (character desc) | Input prompt (background description)<br><br>Prompts are concatenated | Input sample image (character)<br><br>Input sample image background)<br><br>Output image | CLIP - image / DINOv2 cosine similarity score |
| 9hi6li, The image is a medium shot of a young man, presumably a teenager, standing in front of a shelf of books. He has short, dark brown hair and dark brown eyes. He is wearing a white collared shirt. He has a slight smile on his face. | gh1b11, The image depicts a cluster of buildings nestled amongst lush greenery, possibly on a hillside or elevated area. A prominent structure is a two-story building, painted a pale yellow with green accents, situated near the top of the composition; atop it flies a flagpole with two flags, one white and red and one blue and white. Several other buildings, mostly with dark-colored roofing and lighter-colored walls, are interspersed among the trees and foliage. The overall scene is bathed in a warm, golden light, suggesting either sunrise or sunset. | Inputs<br>  <br>  <br>Output<br>  | face CLIP: **0.914**<br>face DINO: **0.565**<br>background CLIP: **0.627**<br>background DINO: **0.350** |
| 9hi6li, The image shows a character with dark green, almost black, straight-cut bob with bangs, standing with pink and purple flowers. The character has fair skin, green eyes, and a gentle smile. They are wearing a white garment with blue trim. | gh1b11, The image is a painted scene of a lush garden leading to a gazebo with an ocean view. A stone wall covered in pink climbing roses is seen on the left, next to a six-sided wooden gazebo with a red-tiled roof and a round attic window. A person wearing a hat can be seen inside the gazebo looking out. A stone path with steps leads from the foreground towards the gazebo and the ocean. The garden is filled with various colorful plants, including reds, whites, purples, and yellows. Overhanging the garden from above are leafy green branches. | Inputs<br>  <br>  | face CLIP: **0.902**<br>face DINO: **0.533**<br>background CLIP: **0.738**<br>background DINO: |

| | | | |
|---|---|---|---|
| | | output  | **0.571** |
| 9hi6li, The image shows an animated young woman standing in what appears to be a rainy environment. She is partially shielded by a large, broad green leaf, possibly from a plant like taro, which she is holding up by its stalk. The leaf has numerous clear water droplets on its surface. The woman has shoulder-length brown hair, with a wave on one side, fair skin, and light-colored eyes. She is wearing a top that is brown around the chest and shoulders, transitioning to a light yellow for the arms and lower portion. | gh1b11, The image is a hand-painted style illustration of a three-story building, presumably an older structure with weathered brown wooden paneling. Each story features multiple windows with blue panes, some with framed decorative molding. A balcony with a stained-glass arched window is central to the second story. The balcony appears to have a weathered grey-painted trim and is supported by two grey pillars. The building is festooned with numerous vertical banners, some hanging crookedly. The banners are white with black Japanese text. Some appear to be protest signs. Laundry is hanging from the windows on the left side. There is greenery at the building's base, indicating it may be a ground-level view. The sky is lightly blue in the upper right corner, with ver | Inputs   Output  | face CLIP: **0.824** face DINO: **0.392** backgr ound CLIP: **0.728** backgr ound DINO: **0.124** |
| 9hi6li, The image is a medium shot of a young character with dark, choppy hair, light brown skin, and wide, light-colored eyes. They have white markings on their cheeks. The character is wearing a brown, fur-like garment over a darker top, secured with a light yellow clasp at the neck. Their arms are covered with green bracers. They are holding a bow in one hand and raising the other hand, gesturing with three fingers extended. | gh1b11, The image depicts a serene and picturesque scene centered around a traditional Japanese house. The house has a dark, tiled roof with defined rows of tiles and a wooden frame that supports the roof's overhang. The house features shoji screens that obscure the view inside. A raised wooden platform serves as a porch, with a stone step leading up to it. The interior has a picture frame and shelves on the back wall. Outside, a garden surrounds the house, filled with round, lush green bushes and trees. A large pine tree, with its characteristic sprawling branches, looms behind the house, partially covered by a cherry tree in | Inputs   Output | face CLIP: **0.898** face DINO: **0.374** backgr ound CLIP: **0.640** backgr ound DINO: |

| | | | |
|---|---|---|---|
| | full bloom, with delicate white and light pink flowers. The sky in the background is a soft gradient of pink and blue hues, suggesting either dawn or dusk. The overall atmosphere is tranquil and peaceful. |  | **0.546** |



CLIP and DINOv2 embeddings for original input face and background images with respect to output for L2, Cosine and Manhattan similarity metrics. Tested on two unique identifier tokens for face and background.

## 3.4 Fréchet Inception Distance (FID) Score

For evaluating our fine-tuned Stable Diffusion model, we use the Fréchet Inception Distance (FID) score, which measures how similar our generated images are to our real target images, with lower scores being better. We achieved an FID of 208.819. This score is generally considered high; **for context, FID scores closer to 0 represent better similarity, with scores in the single digits or low tens often indicating high-quality generation on standard benchmarks, though the definition of a "good" score can be relative to the dataset's complexity.** This suggests a significant difference between our generated images and our target distribution. However, the relevance of this specific score is questionable for us because we fine-tuned our model using fast-DreamBooth, which inherently involves a very small number of training images. More critically, we then calculated this FID score using only 14 images each from our dataset and 14 generated ones, a number vastly insufficient for a trustworthy result.

The FID score's trustworthiness hinges on accurately estimating the mean and, crucially, the covariance matrix of high-dimensional features (2048-dimensional from the InceptionV3 model our code uses) extracted from large image sets.**Typically, a minimum of 10,000 samples from each distribution (real and generated) is recommended for a stable and reliable FID score, with many benchmark evaluations using 30,000 to 50,000 images to ensure robust statistical estimates.** With just 14 samples, these estimates, particularly for the covariance, are inherently unstable and unlikely to represent the true distributions. While our calculation code employed a common stabilization technique by adding a small numerical offset to the covariance matrices, this primarily prevents computational errors and doesn't make the resulting

FID score meaningful or a fair assessment of our model's quality given the severe lack of data. The general FID calculation process involves passing real and generated images through an InceptionV3 model to get feature activations, modeling these activations as multivariate Gaussian distributions for each set, and then computing the Fréchet distance between these distributions using their means (μr, μg) and covariance matrices (Σr, Σq) via the formula:

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

# 4. Discussion and conclusions

In this section we are going to discuss our key findings from various prompts and configurations.

## 4.1 Ablation Studies

In this section we evaluate multiple experimental knobs. Unless explicitly specified we use the following prompt.

“gh1b11, The image depicts a charming, overgrown cottage nestled in a lush, green landscape. The cottage is the central focus, with its walls and roof almost entirely covered in vibrant green vines and moss. Windows with white frames are visible, some with awnings, and one with an open shutter revealing a pink curtain. A girl in a light grey dress and a red bow runs towards the dark doorway of the cottage. A stone pathway leads from the cottage into a colorful garden filled with flowers of various colors, including pink, yellow, and purple. Tall trees surround the cottage, creating a sense of seclusion and tranquility. To the right, there is a secondary building with a rounded doorway, partially obscured by birch trees. The overall atmosphere is whimsical and idyllic, with a focus on nature and a sense of enchantment”

This prompt is present in our fine-tuning dataset and represents the following image:

The model we use in these experiments is the “ghibli-background-long” model, which is better suited for generating background images than subjects. As a result, we don’t expect the model to do a great job at generating the girl and will not be using it as a metric.

### 4.1.1 Sampler Choice

We have an option to select one of many samplers to use during the diffusion process. We have a lot of options. In this section, we display the ones we found interesting and compare their characteristics.

Euler A

DPM++2M Karras

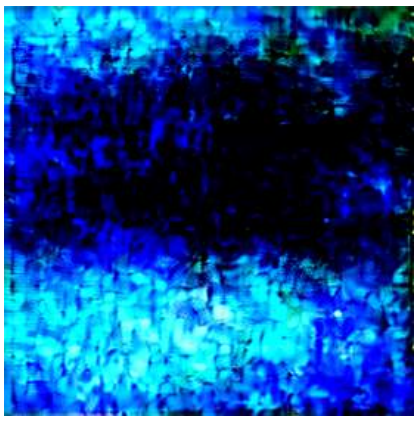
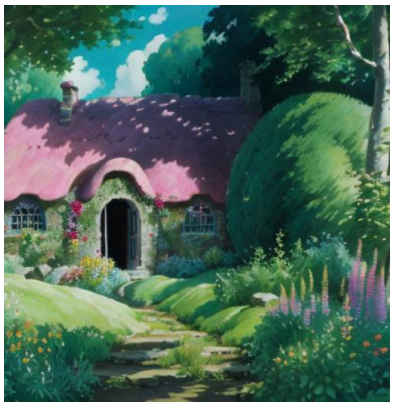
LMS

DDIM

Qualitatively, We find “Euler A” to be the best sampler for generation quality and adherence to the prompt for a wide variety of prompts. DDIM also displays a good understanding of the prompt, and appears to have a preference for bright, pastel colors. DPM++ 2M Karras was unable to center the cottage and uses a more muted color palette. LMS also showcases a limited text embedding quality as it generated 2 houses. This might indicate that LMS does not retain spatial information to the same degree as other samplers.

### 4.1.2 Number of Sampling Steps

For Euler A sampler, we find that performing diffusion to be done over more sampling steps shows a marked improvement in generation quality.

<table>
<tr>
<td>  <br>Generation Steps = 1</td>
<td>  <br>Generation Steps = 50</td>
</tr>
<tr>
<td>  <br>Generation Steps = 100</td>
<td>  <br>Generation Steps = 150</td>
</tr>
</table>

One interesting observation we note is that 100 generation steps generates a girl entering the house, as expected. However, we found this to not be the case at 150 generation steps. Since our model is trained on background images, our understanding is that the sampling process begins to treat the subject as extra noise in its goal to generate a good background.

### 4.1.3 Keyword Specificity

All of our captions in the dataset begin with the unique identifier “gh1b11” on suggestion of the DreamBooth paper. We observe that dropping this identifier at test time leads to a marked drop in quality.

| | |
|---|---|
|  Without gh1b11 |  With gh1b11 |

Dropping the keyword still generates an image that adheres to the prompt but the image does not seem to be significantly in the style of a ghibli image.

### 4.1.4 Number of Epochs

The image below shows the improvement in image generation from left to right on a prompt with respect to epochs 500, 1500, 2500, 3500 respectively. The model improved with epochs and we saw better reconstruction of eyes in the image.

| Prompt | Epochs 500, 1500, 2500, 3500 (top to bottom) |
|---|---|
| A stylized anime illustration depicting a young girl walking alongside a vast, vividly detailed green leaf that dwarfs her figure, emphasizing a miniature or shrunken perspective. The girl, dressed in a deep red outfit with a cream-colored sleeve and a khaki strap crossing her chest, wears her auburn hair in a practical ponytail, fastened with what appears to be an oversized clothespin-like accessory, hinting at a whimsical or fantastical world where ordinary objects are reimagined as tools. Her downward gaze and solemn expression convey a sense of quiet determination or introspection. The lush, veiny leaf in the background is rendered with painterly strokes and subtle tonal variations of green, punctuated by occasional brown blemishes and illuminated by soft, diffuse light. The |   |

| | |
|---|---|
| girl's shadow, cast crisply against the leaf, mirrors her posture and adds a grounding sense of realism. The overall composition evokes a serene yet adventurous atmosphere, blending natural wonder with a touch of magic. |  <br>  |

### 4.1.5 Training StableDiffusion 1.5

While our best results were obtained with the StableDiffusion 2.0 model, we initially started by fine-tuning StableDiffusion 1.5. Here are the results from it:

| | |
|---|---|
|  |  |
|  |  |

While this fine-tuned model is still very good at generating backgrounds, we find it's quality in generating subjects, such as human faces and cats to be much worse.

### 4.1.6 Overfitting test

For our training dataset, we develop a duplicate dataset with the same images but rephrased captions. We use this dataset to check whether our model has overfit our text captions.

| | |
|---|---|
|  Prompt with rephrased caption |  Prompt with rephrased caption |

Although these images are stylistically different, both of them are qualitatively good images that adhere to the prompt. We conclude that our model has learned visual and text representations well.

### 4.1.7 Classifier-Free Guidance (CFG) Scale Modulation

The CFG scale (classifier-free guidance scale) or guidance scale is a parameter that controls how much the image generation process follows the text prompt. The higher the value, the more the image sticks to a given text input.

| 1 | 10 | 12.5 | 15 | 30 |
|---|---|---|---|---|
|  |  |  |  |  |

Prompt: A stylized anime illustration depicting a young girl walking alongside a vast, vividly detailed green leaf that dwarfs her figure, emphasizing a miniature or shrunken perspective. The girl, dressed in a deep red outfit with a cream-colored sleeve and a khaki strap crossing her chest, wears her auburn hair in a practical ponytail, fastened with what appears to be an oversized clothespin-like accessory, hinting at a whimsical or fantastical world where ordinary objects are reimagined as tools. Her downward gaze and solemn expression convey a sense of quiet determination or introspection. The lush, veiny leaf in the background is rendered with painterly strokes and subtle tonal variations of green, punctuated by occasional brown blemishes and illuminated by soft, diffuse light. The girl's shadow, cast crisply against the leaf, mirrors her posture and adds a grounding sense of realism. The overall composition evokes a serene yet adventurous atmosphere, blending natural wonder with a touch of magic.

Increasing the CFG (Classifier-Free Guidance) value in AI image generation beyond an optimal range (typically 7-12, with issues often arising above 15-20) can lead to pixelated images due to several interconnected factors. Over-adherence to the text prompt restricts the AI's creative freedom, causing it to produce distorted, unnatural results with a loss of fine detail and resolution. This extreme adherence can also introduce various visual artifacts, excessive

contrast, and over-saturation, all of which contribute to a degraded, coarse, or pixelated appearance as the image's coherence and subtlety break down. Essentially, while aiming for stricter prompt following, very high CFG values sacrifice overall image quality, pushing the generation process into a state where visual fidelity is compromised.

### 4.1.8 Increasing Prompt Complexity

We steadily increased the complexity of the prompt to see how the complexity is related to the generated results.
Base prompt: A stylized anime illustration depicting a young girl walking alongside a vast, vividly detailed green leaf that dwarfs her figure, emphasizing a miniature or shrunken perspective. The girl, dressed in a deep red outfit with a cream-colored sleeve and a khaki strap crossing her chest, wears her auburn hair in a practical ponytail, fastened with what appears to be an oversized clothespin-like accessory, hinting at a whimsical or fantastical world where ordinary objects are reimagined as tools. Her downward gaze and solemn expression convey a sense of quiet determination or introspection. The lush, veiny leaf in the background is rendered with painterly strokes and subtle tonal variations of green, punctuated by occasional brown blemishes and illuminated by soft, diffuse light. The girl's shadow, cast crisply against the leaf, mirrors her posture and adds a grounding sense of realism. The overall composition evokes a serene yet adventurous atmosphere, blending natural wonder with a touch of magic.

| Modified Prompt | Result |
|---|---|
| A stylized anime scene of a solemn girl beside a giant leaf, blending magical realism with a miniature perspective. |  |
| This is a stylized anime illustration depicting a young girl walking alongside a vast, vividly detailed green leaf that dwarfs her figure, emphasizing a miniature or shrunken perspective. The girl, dressed in a deep red outfit with subtle stitching and a cream-colored sleeve partially rolled up, sports a khaki strap crossing her chest that appears to support a satchel. Her auburn hair is tied in a practical ponytail, fastened with what resembles an oversized wooden clothespin-like accessory, suggesting a whimsical or fantastical world where everyday items are creatively reimagined as tools. Her downward gaze and solemn expression convey quiet determination or introspection. The lush, veiny leaf in the background is rendered with painterly strokes, intricate venation, and subtle tonal variations of green, punctuated by occasional brown blemishes and softly illuminated by diffused sunlight filtering through unseen foliage. The girl's shadow, cast crisply against the leaf’s textured surface, mirrors her posture and adds a grounding sense of realism. The overall composition evokes a serene yet adventurous atmosphere, blending natural wonder with a quiet sense of magical realism. |  |

| | |
|---|---|
| This is a stylized anime illustration portraying a diminutive young girl progressing alongside an enormous, intricately rendered green leaf whose sheer scale diminishes her presence, accentuating a miniature or shrunken visual narrative. The girl, attired in a deep crimson ensemble embellished with fine seamwork and paired with a cream-colored sleeve that subtly wrinkles at the elbow, is adorned with a diagonally slung khaki strap, suggestive of utilitarian function, perhaps anchoring a concealed pouch or gear. Her auburn hair is meticulously gathered into a no-nonsense ponytail, secured by a disproportionately large, clothespin-like contraption—an imaginative recontextualization of mundane objects into fantastical accoutrements, reinforcing the surreal logic of the world she inhabits. Her downward-cast eyes and grave expression suggest a layered emotional state, poised between contemplative resolve and quiet sorrow. The immense leaf, textured with painterly veining and chromatic gradients of viridian, jade, and subtle ochres, is sporadically marked with blemishes that enhance its organic realism, all softly bathed in ambient, filtered sunlight that imparts a tranquil glow. The girl's sharply defined shadow, aligned precisely with the curvature of the leaf, anchors her within the scene, imbuing the composition with spatial coherence and physical believability. The overall mise-en-scène conjures a delicate tension between serenity and latent adventure, merging botanical grandeur with the intimate surrealism of a world both familiar and impossibly reimagined. |  |
| This is a meticulously stylized anime illustration depicting a seemingly inconspicuous yet narratively charged young girl traversing the longitudinal expanse of an immense, hyper-realistically articulated green leaf whose monumental scale renders her almost imperceptible, evoking a profound sense of disproportion that gestures toward a deliberate manipulation of spatial hierarchies and a metaphorical exploration of perspective and vulnerability. The girl is clad in a richly hued crimson garment, tailored with ornamental stitching that subtly references traditional craftsmanship, and complemented by a cream-toned sleeve whose creases and fabric tension suggest motion and tactile authenticity; a khaki strap bisects her torso, possibly affixed to a concealed satchel or tool-bearing harness, implying utilitarian preparedness within a semi-domestic yet alien context. Her auburn hair, drawn into a tightly bound ponytail, is clasped with a surreal, oversized clothespin-like device—an object both absurd and functional—imbuing the scene with a sense of diegetic whimsy and suggesting a universe governed by an alternate taxonomy of scale and symbolism in which everyday artifacts assume fantastical significance. Her downturned gaze, combined with a subtly furrowed brow and melancholic poise, communicates an interiority rich with contemplative gravity, perhaps suggestive of a larger, unseen journey or internal narrative arc. The leaf beneath and behind her unfurls in sweeping painterly gestures, its surface intricately layered with variegated tones of chlorophyll-rich green, moss, and intermittent earthen blemishes that add botanical verisimilitude, while dappled, diffuse sunlight penetrates an implied canopy, creating an interplay of light and shadow that enhances the diorama’s textural complexity. A crisply delineated shadow of the girl mirrors her precise stance with uncanny fidelity, anchoring her figure to the terrain and reinforcing the illusionistic cohesion of the composition. Altogether, the image synthesizes elements of quiet introspection, environmental monumentality, and latent magical realism into a tableau that is at once intimate and mythic, evoking a world where nature’s minutiae are magnified into epic significance and the mundane is rendered transcendently strange. |  |

We observed that while adding complexity to a prompt initially refines the output, exceeding a certain threshold of intricacy can paradoxically lead to a decline in the quality of the generated images. The phenomenon where increasing prompt complexity in AI image generation initially improves results but then leads to deterioration at a certain point can be attributed to how these models process and interpret information. While detailed prompts generally provide more

guidance and lead to more accurate and relevant outputs, overly long or complex prompts can confuse the AI. These models have limitations in how much information they can effectively handle; words at the beginning of a prompt are often prioritized more than those at the end, meaning longer prompts may cause the AI to overlook or miss later instructions.
Over-complicating a prompt with too much technical jargon, excessively complex sentences, or multiple distinct ideas packed into one can overwhelm the AI, leading to outputs that are off-target, convoluted, or cluttered with key elements lost. Essentially, there's a sweet spot for prompt complexity; providing enough detail is crucial, but exceeding the AI's capacity to cohesively synthesize all the information can result in a decline in the quality and coherence of the generated image.

## 4.2 Color Problem

By improving our captioning and training process, we were able to solve the problem that we proposed in Section 2.2. Our model is able to swap out colors that are different from the dataset without any loss in generation quality.

| | |
|---|---|
|  (...) She wears a deep red outfit with a cream sleeve and khaki strap, her auburn hair tied in a ponytail. (...) |  (...) She wears a deep blue outfit with a cream sleeve and khaki strap, her golden hair tied in a ponytail (...) |

# Statement of individual contribution

All group members collaborated closely throughout the project. We regularly held meetings, discussed ideas, and worked together on implementation and evaluation. While responsibilities often overlapped, each member focused more heavily on specific components:

1. Aryaman Nasare – Led dataset curation, including manual image selection and caption refinement across all subsets.
2. Ridhwik Kalgaonkar – Focused on evaluation methodology, including CLIP/DINOv2 analysis and qualitative inspection.
3. Aditya Potnis – Primarily handled model training pipelines, experimental runs, and integration of tools like fastDreamBooth.

4. Kartik Ramesh – Drove ablation studies, hyperparameter tuning, and experimental design.

# References

## Papers

[1] https://arxiv.org/abs/2208.12242 (Dreambooth:Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation by Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman)
[2] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," Feb. 02, 2024, arXiv: arXiv:2304.07193. doi: 10.48550/arXiv.2304.07193.
[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

## External Code, Datasets

- Studio Ghibli Movies: https://huggingface.co/datasets/Nechintosh/ghibli
- Dataset of paintings by John Singer Sargent (WikiArt): https://drive.google.com/file/d/1capT9kF-zCu2OiNVzm7VG5DQDaAQLI1Q/view
- Dreambooth Paper:
- Sample implementation on Stable diffusion 1.5: https://colab.research.google.com/github/TheLastBen/fast-stable-diffusion/blob/main/fast-DreamBooth.ipynb
- OpenAI CLIP: https://openai.com/index/clip/
- Meta DINOv2: https://dinov2.metademolab.com/
- CFG Scale: https://getimg.ai/guides/interactive-guide-to-stable-diffusion-guidance-scale-parameter
- Prompt Complexity: https://openart.ai/blog/post/the-most-complete-guide-to-stable-diffusion-parameters
- amulbel/ghibli-movies-dataset: https://www.kaggle.com/datasets/amulbel/ghibli-movies-pictures/data
- Datasets - All of our curated datasets (needs illinois account to access)
- cs444diffusion - Our trained models and datasets (needs illinois account to access)