Applying ControlNet to Custom Structural and Contextual Conditioning Tasks

Team ID: 14

Akash Gorakh Chaudhari MT2024012 **Shardul Sisodiya** MT2024140

16 December 2025

Project Repository: https://github.com/shardul523/Control-Net-Project

Abstract

Diffusion-based generative models such as Stable Diffusion have demonstrated impressive image synthesis capabilities, yet they lack fine-grained controllability over structural layout and semantic emphasis. **ControlNet** addresses this limitation by introducing external conditioning signals that guide the diffusion process while preserving the expressive power of large pretrained models.

In this project, we investigate ControlNet across two complementary controllable image generation tasks: (1) **Anime Sketch Coloring**, where line-art sketches act as a hard structural constraint, and (2) **Saliency-Guided Generation**, a novel task where contextual saliency masks provide a soft spatial prior. We design custom conditioning representations, construct paired datasets, and train ControlNet models using the Hugging Face `diffusers` library. Through systematic ablation studies on learning rate, conditioning strength, and training duration, we demonstrate that ControlNet generalizes beyond purely geometric constraints, enabling controllable generation guided by both structural and contextual signals.

1 Introduction

Text-to-image diffusion models have emerged as a powerful paradigm for high-quality image synthesis. However, reliance on text prompts alone offers limited guarantees regarding spatial structure, object placement, or compositional consistency. This lack of explicit control poses challenges for applications that require predictable and constrained generation.

ControlNet introduces an architectural extension that preserves the pretrained diffusion backbone while enabling external conditioning through trainable control layers. By injecting control signals at multiple stages of the denoising process, ControlNet enforces adherence to structural or contextual constraints without degrading the base model's generative capabilities.

In this project, we explore ControlNet under two complementary paradigms:

- **Hard Structural Control**, where the conditioning signal encodes explicit geometric boundaries (anime line art).
- **Soft Contextual Control**, where the conditioning signal encodes spatial importance rather than precise geometry (saliency masks).

By implementing both paradigms, we satisfy the project's requirement for novelty while demonstrating technical rigor and a deep understanding of controllable diffusion models.

2 Related Work

Stable Diffusion operates in a latent space to enable efficient high-resolution image synthesis and serves as the frozen backbone for all experiments in this project.

ControlNet extends Stable Diffusion by cloning and freezing the original encoder blocks while introducing trainable control layers initialized to zero. This design allows the model to learn new conditioning modalities without catastrophic forgetting of pretrained knowledge.

Prior work on ControlNet has largely focused on geometric conditioning signals such as edges, depth maps, poses, and segmentation masks. In contrast, this project explores **contextual saliency masks** as a novel control modality. Unlike geometric constraints, saliency maps encode semantic importance rather than exact structure, enabling a softer and more flexible form of controllable generation.

3 Dataset Creation Pipeline

We constructed two datasets consisting of aligned triplets: *(Target Image, Conditioning Image, Text Prompt)*. A unified preprocessing pipeline was implemented to ensure spatial and semantic alignment.

3.1 Task 1: Anime Sketch Coloring

- **Source Dataset:** Naruto BLIP Captions Dataset
- **Target Images:** Colored anime illustrations
- **Control Signal:** Anime-style line-art sketches extracted from the original colored images. These sketches capture the prominent facial contours, hair outlines, and structural edges of the characters while omitting color and texture information. The line-art representations closely resemble hand-drawn anime sketches and provide strong structural guidance for the coloring task.
- **Extraction Method:** The control signals are generated using the `LineartAnimeDetector` from the `ControlNet` auxiliary library. This detector utilizes an XDoG (Extended Difference of Gaussians)-based edge extraction technique, which is well suited for anime images. XDoG enhances sharp edges and smooth contours, producing clean and high-quality sketches that preserve facial structure and stylistic details typical of anime illustrations.
- **Final Dataset Link:** anime_faces
- **Dataset Size:** 1220 samples

3.2 Task 2: Saliency-Guided Generation

- **Source Dataset:** SALICON: Saliency in Context Dataset
- **Target Images:** Natural RGB images

- **Control Signal:** Contextual saliency masks. Unlike the rigid geometric boundaries of line art, these masks represent regions of visual importance, where pixel intensity corresponds to human attention. Brighter regions indicate the primary subject or focal point, while darker regions represent background context. This serves as a "soft" spatial prior, guiding the model on *where* to generate high-detail content without strictly constraining the internal shapes or textures, allowing for greater semantic flexibility during generation.
- **Extraction Method:** The SALICON dataset provides pre-computed human attention–based saliency maps, derived from large-scale mouse-tracking and eye-tracking data. These saliency masks are directly used as control inputs without requiring additional saliency model inference. BLIP2 has been used for generating the captions for images to be used as control signals.
- **Final Dataset Link:** final_dataset
- **Dataset Size:** 10000 samples

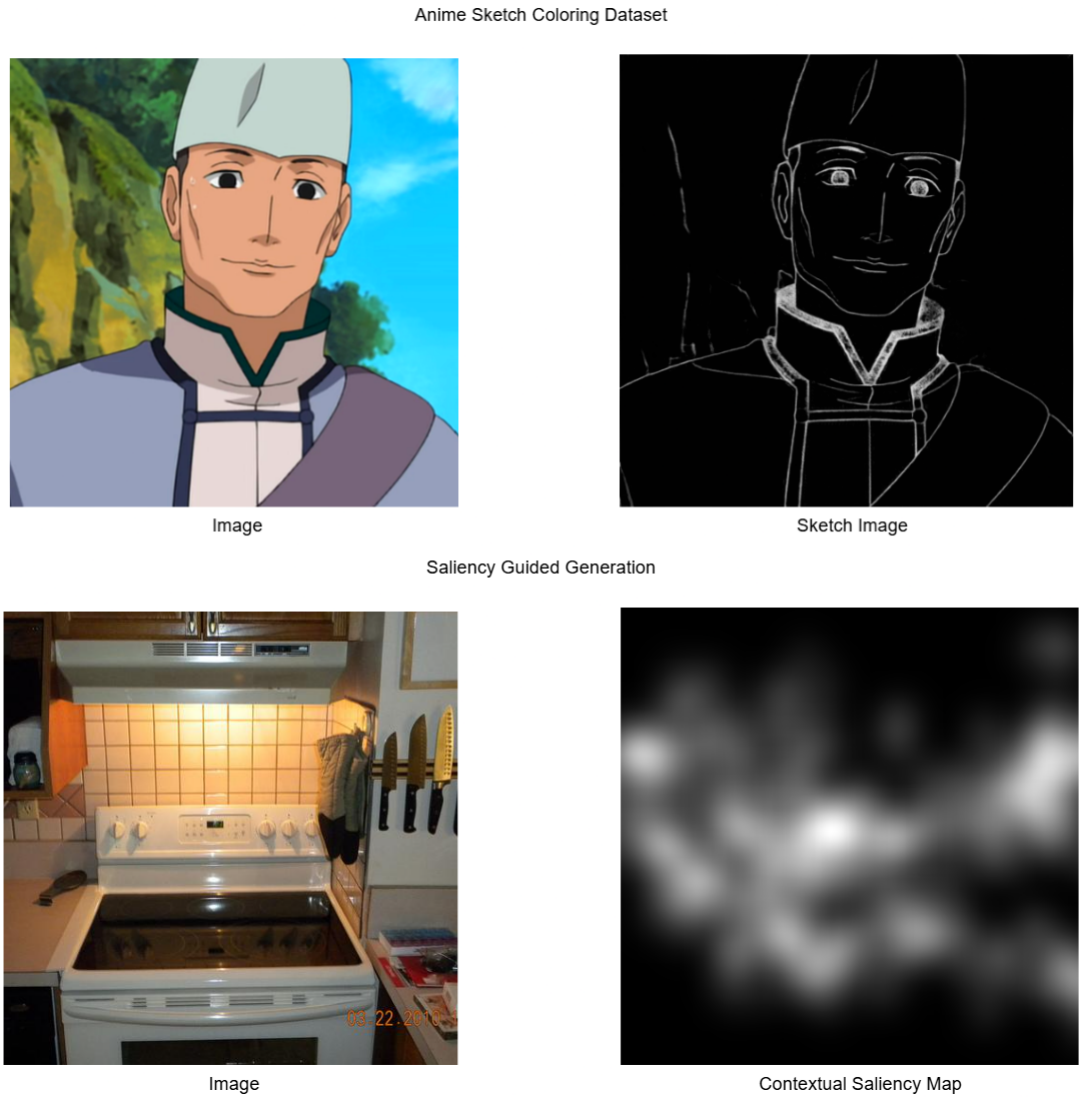

Figure 1: Data samples from both tasks. Top: Anime image and extracted line art. Bottom: Natural image and extracted saliency mask.

4 Model Architecture and Training Setup

4.1 Architecture

4.1.1 Anime Sketch Coloring Task

- **Framework:** Hugging Face `diffusers`
- **Training Script:** Custom ControlNet training loop
- **Base Diffusion Model:** `stablediffusionapi/anything-v3`
- **Text Encoder and Tokenizer:** `runwayml/stable-diffusion-v1-5`
- **Trainable Component:** ControlNet only (UNet, VAE, and text encoder frozen)

4.1.2 Saliency Guided Generation

- **Framework:** Hugging Face `diffusers`
- **Training Script:** Modified `train_controlnet.py`
- **Base Model:** `runwayml/stable-diffusion-v1-5`

To improve training efficiency, gradient checkpointing and mixed-precision training (bfloat16) were enabled using the Accelerate framework. Optimization was performed using the 8-bit AdamW optimizer, ensuring stable convergence while maintaining memory efficiency.

4.2 Training Configuration

Hyperparameter	Anime Task	Saliency Task
Batch Size	24	32
Learning Rate	2×10^{-5}	1×10^{-5}
Optimizer	AdamW (8-bit)	AdamW (8-bit)
Max Steps / Epochs	100 epochs	5000 steps

Table 1: Training hyperparameters for both tasks.

4.3 Training Loss Curve

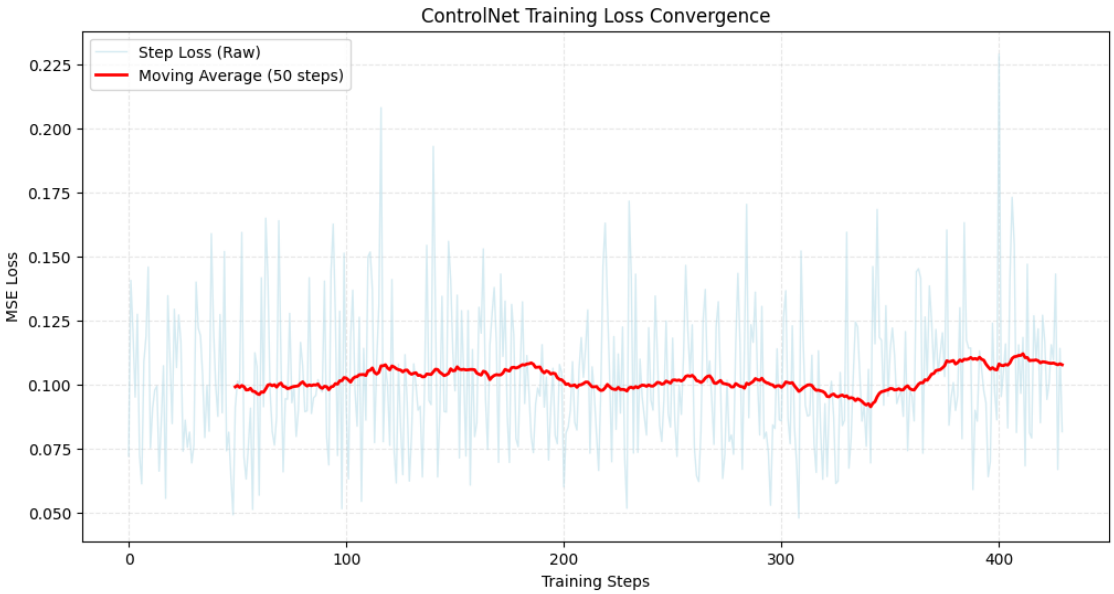

Figure 2: **Anime Sketch Coloring Training Curve.**During training, the loss was logged every 10 steps, resulting in approximately 450 recorded values for 4500 training steps. The x-axis in the plotted training loss curve was incorrectly represented; the correct training steps should be obtained by multiplying the x-axis values by 10.

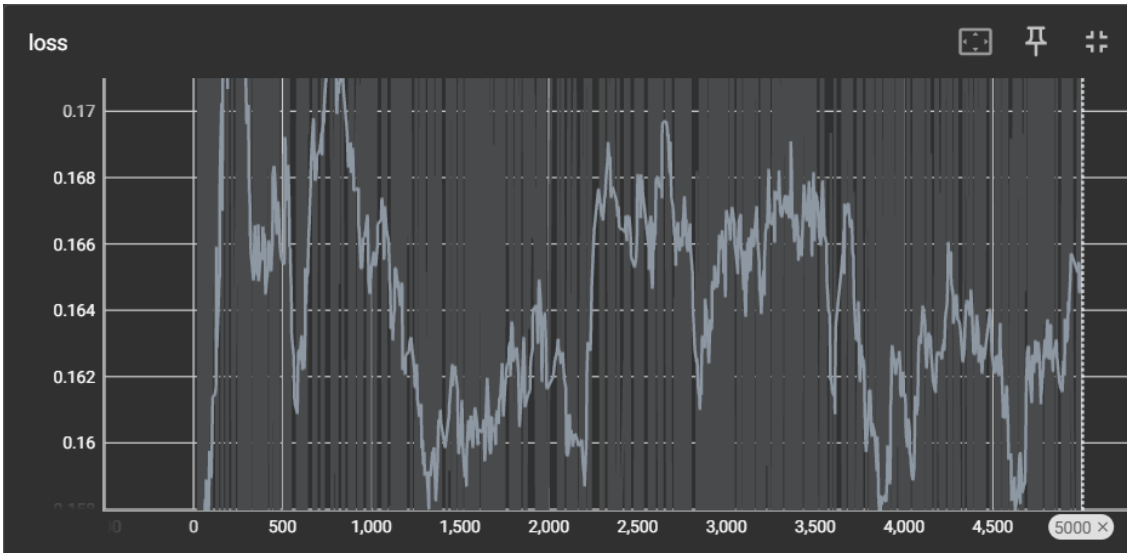

Figure 3: **Saliency Guided Generation Training Curve**. The curve was smoothed using an `exponential moving average` with a decay factor of `0.98` to reduce high-frequency noise while preserving meaningful changes in training dynamics

5 Experiments and Ablation Studies

5.1 Conditioning Scale Ablation

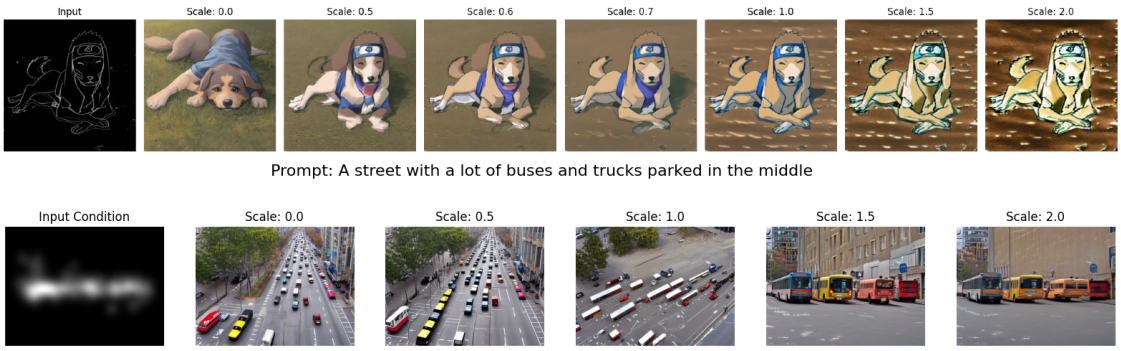

Figure 4: As the conditioning scale increases, the images generated by the model are more heavily influenced by the conditioning image.

5.2 Training Duration Ablation

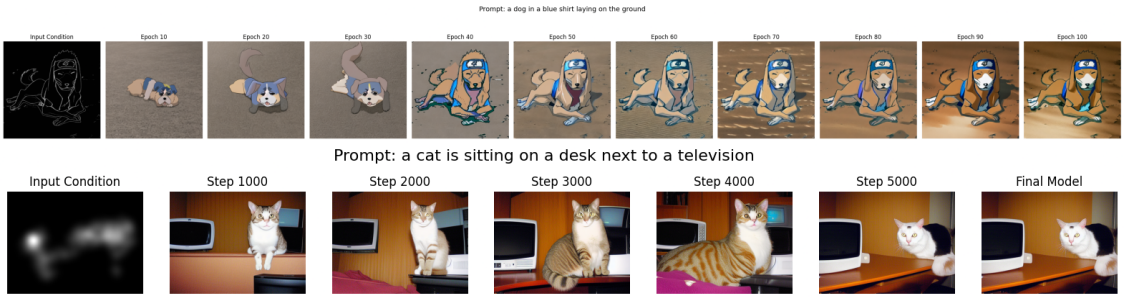

Figure 5: As the training duration increases, the model learns to follow the conditioning signal being provided.

5.3 Prompt Ablation

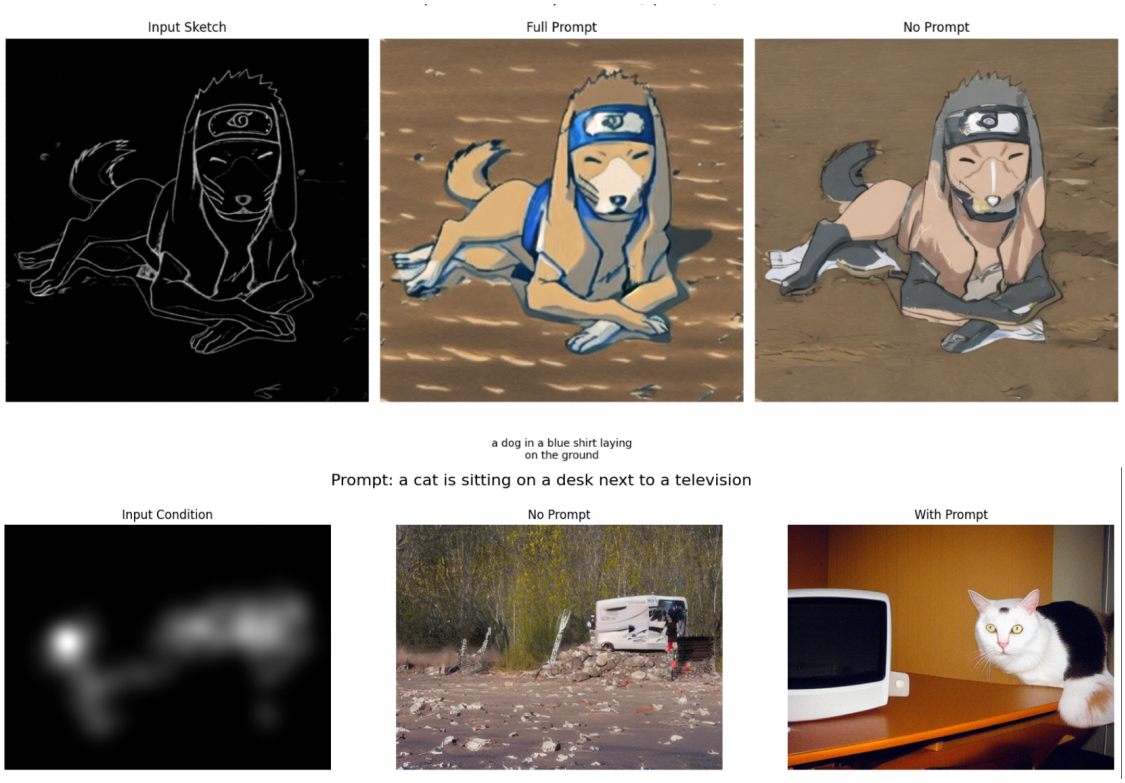

Figure 6:
Even without the prompt, the anime sketch coloring can produce legible results since the conditioning image provides the structure to be followed.
But with contextual saliency mask, which highlights only the region where the image's context should be present, it was upto the text prompt to guide the model. Without it, the model generates random gibberish.

6 Evaluation

6.1 Qualitative Evaluation

The anime model successfully colorized sketches while preserving line structure. The saliency-conditioned model demonstrated effective soft control, generating semantically appropriate objects aligned with the saliency prior.

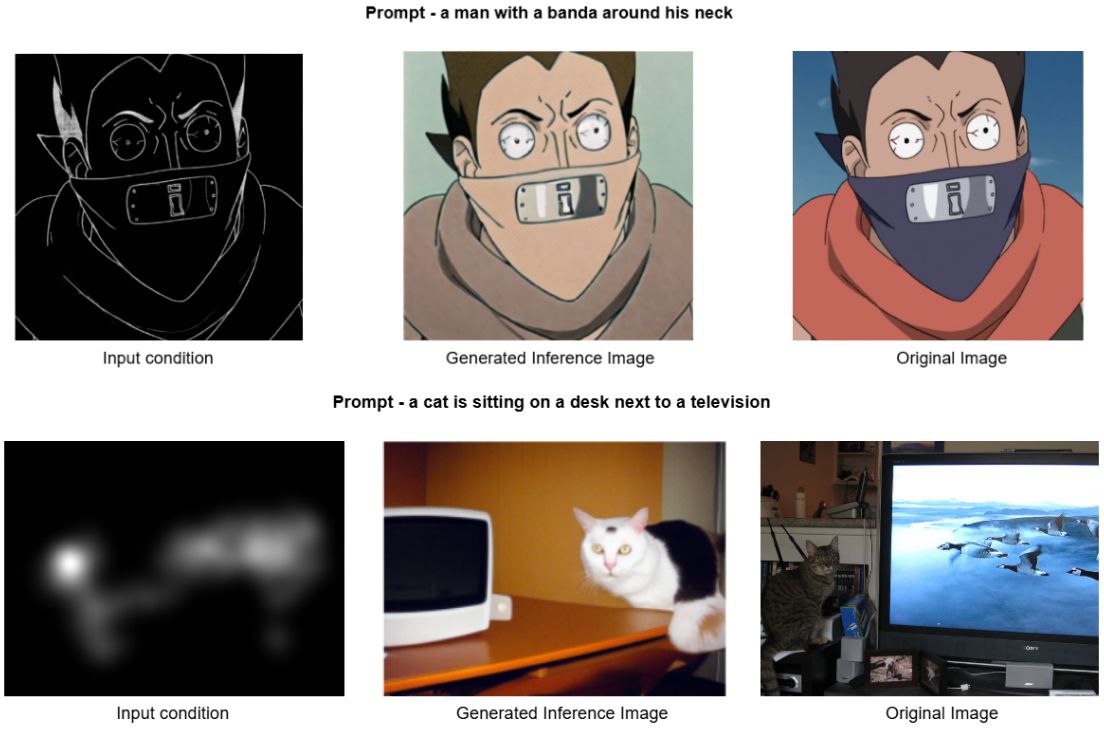

Figure 7: Qualitative comparison. Left: Input condition. Middle: Generated output. Right: Ground truth.

6.2 Quantitative Evaluation

Model Variant	SSIM (↑)	LPIPS (↓)	CLIP Score (↑)
Anime Task (Final)	0.5549	0.5762	-
Saliency Task (Low LR)	0.1975	0.8022	32.1987
Saliency Task (Standard LR)	0.2008	0.7944	32.1636
Saliency Task (High LR)	0.1949	0.7822	32.1551

Table 2: Quantitative evaluation results.

6.2.1 Discrepancy Between Quantitative Metrics and Visual Quality

Although the quantitative metrics in `Table 2` suggest relatively low `SSIM` and high `LPIPS` values—particularly for the saliency-guided task—the qualitative results demonstrate visually coherent and semantically meaningful generations. This apparent discrepancy arises from a fundamental

mismatch between traditional pixel- or patch-based similarity metrics and the objective of conditional generative modeling.

SSIM and LPIPS measure similarity to a specific ground-truth image, penalizing deviations in texture, color distribution, and fine-grained spatial details. However, ControlNet is explicitly designed to allow appearance diversity while enforcing adherence to a conditioning signal. As a result, the generated images often differ substantially from the original target image at the pixel level while still satisfying the structural or contextual constraints imposed by the control signal.

This effect is particularly pronounced in the saliency-guided generation task, where the conditioning signal encodes only a soft spatial prior rather than rigid geometry. In this setting, multiple visually plausible outputs may exist for a single saliency mask and prompt, making strict pixel-level correspondence both unnecessary and undesirable. Consequently, lower SSIM and higher LPIPS scores do not indicate poor generative quality, but rather reflect the model's ability to generate diverse yet semantically aligned outputs.

The qualitative comparisons (Figure 7) and CLIP similarity scores further support this interpretation, demonstrating that the model maintains semantic alignment with the prompt and conditioning signal despite reduced pixel-level similarity to the ground truth.

7 Conclusion

This project demonstrates ControlNet's versatility across both hard structural and soft contextual conditioning regimes. Line-art sketches provide precise spatial constraints, while saliency masks introduce a novel attention-based control mechanism. Through systematic ablations and careful tuning, we establish a robust and reproducible ControlNet training pipeline.

8 Demo Application

To demonstrate the practical usability of the trained ControlNet models, we implemented an interactive demo using a Jupyter notebook interface. The demo allows users to provide a custom conditioning image along with a textual prompt and generates a corresponding output image using the trained ControlNet.

The demo supports both tasks explored in this project:

- **Anime Sketch Coloring:** Users upload a line-art sketch and optionally provide a prompt describing color or style.
- **Saliency-Guided Generation:** Users upload a saliency mask and provide a semantic prompt to guide the generated content.

For the purpose of running the demo, we have provided saliency masks within the repository since the information regarding how these saliency masks were produced is not clear.

The demo performs preprocessing, ControlNet-based inference, and visualization within a single interface, enabling rapid qualitative evaluation of model behavior under different conditioning signals and prompts. This interactive component highlights the controllability and flexibility of the proposed system and serves as a reproducible demonstration of the project outcomes.