Image Colourization using Diffusion with Multimodal Capabilities

Aditya Kumar[1,2], Ayush Naique[1,2], Divyanshu Sharma[1,2], and Gokulkrishna M[1,2]

[1]New York University, New York
[2]*These authors contributed equally to this work*

May 15, 2024

Abstract

Diffusion models possess strong capabilities for solving image-based self-learning tasks. Image colourization helps such self-learning models possibly generalize the color scheme, which might also help improve performance on object classification, object detection and other downstream tasks. In this project, we try to implement image colorization using diffusion models as well as try to imbue it with multi-modal capabilities. Most of the prior models in this domain use attention and auto-encoding mechanisms. Here, we have explored further architectural changes and also have tried to implement further multi-modal capabilities other than existing ones.

(Code hosted on GitHub: REPO, Hugging Face Spaces: GRADIO)

1 Introduction

Image colorization using diffusion models represents an intriguing intersection of computer vision and artificial intelligence, presenting exciting prospects for elevating visual content. This technique involves enriching grayscale images with vibrant colors using sophisticated algorithms grounded in statistical mechanics and probabilistic modeling.

Fundamentally, image colorization with diffusion models operates by iteratively propagating color information throughout the image space. Leveraging intricate mathematical frameworks like partial differential equations and stochastic processes, these models simulate the diffusion of color from known regions to unknown areas. Through this process, the color distribution gradually evolves, resulting in a coherent and visually appealing outcome.

One of the primary advantages of diffusion-based colorization lies in its ability to autonomously infer colors for grayscale images, eliminating the need for manual intervention. By scrutinizing patterns and structures within the image, these models can make informed decisions regarding color assignments, capturing subtle nuances and details that conventional methods may overlook.

In this document, we thoroughly investigate the fundamental concepts of image colorization using diffusion models. Our primary reference throughout this study is [ZRA23], and we aim to enhance their discoveries while incorporating innovative multimodal capabilities into the image colorization model.

2 Literature Survey

The problem of colorization of images fall mainly under two categories: Colorization of images using Convolutional Neural Networks and Diffusion based approach. Both these approaches have its strength and weaknesses. In Richard Zhan et al [ZIE16] paper colourful image generation using CNN, they tackle the problem of hallucination of color images from greyscale images. The methodology embraces the underlying uncertainty of the problem by posing it as a classification task and use class-re balancing at training time to increase the diversity of colors in the result. The system is implemented as a feed-forward pass in a CNN at test time and is trained on over a million color images. The algorithm is evaluated using a "colorization Turing test", asking human participants to choose between a generated and ground truth color image. They utilize self supervised learning to train a large ConvNet and sample a & b channels to colourize the images.

For Diffusion based approaches, the goal is to leverage a large pretrained generative Diffusion Model, and finetune it for the colorization task without losing its generative power or attention to text prompts [ZAG+23]. This approach of using a large pretrained diffusion modal allows us to colourize the image in a custom way and we can even imbue it with multi-modality [LLZ+24] by using architectures like Control Net [ZRA23]. In this approach, they freeze the pre-trained diffusion model (Stable diffusion [RBL+22]) and introduce zero conv layers to add custom controls. Since the control input is translated to latent space after encoding, this enables us to have any input form (strokes, text and audio) that can be converted to the same latent space. The traditional non-deep learning based approaches use scribbling or transferring color between images, which we do not explore here as we focus on deep learning approaches.

3 Methodology

The project began with a base architecture to facilitate the development of multi-modal capabilities. The Control-Net architecture was employed, which incorporates trainable copies of weights from the state-of-the-art Stable Diffusion model. These trainable weights are integrated with the original fixed weights through zero convolutional layers, as illustrated in Figures [1] and [2].

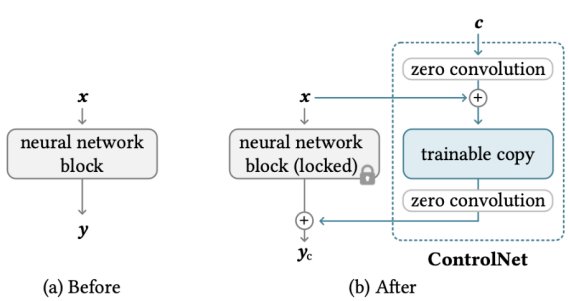


Fig. 1: Zero Convolutional Layer

For unconditional image colorization, the Control-Net model was fine-tuned on a sample of 15K images from the Imagenet database. The model not only colorized images richly but also responded well to prompts such as "color image in pink" despite not being explicitly fine-tuned with such prompts. This indicated that the Stable Diffusion model is highly receptive to input prompts. Further experiments on the input size of

the images revealed that, since Stable Diffusion was trained on images of size 512x512, it does not colorize images as effectively for other dimensions.

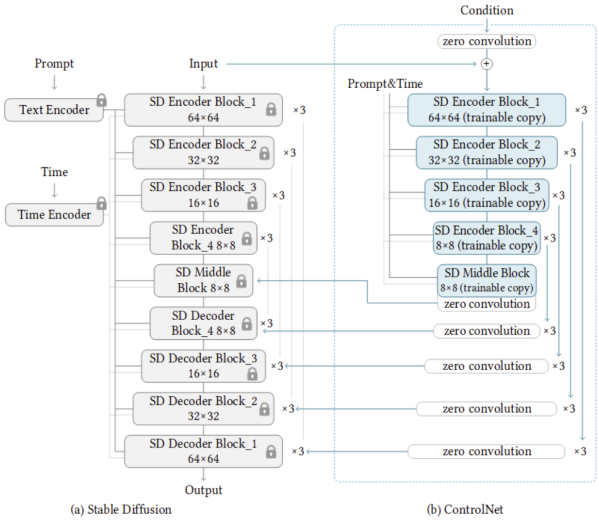


Fig. 2: Control Net Architecture

For conditional colorization, the first objective was to incorporate text prompts specifying the desired coloration of images. The model was trained with prompts using the BLIP model [LLXH22]. The BLIP model generates prompts by leveraging a pre-trained transformer architecture designed for multi-modal tasks, which involves understanding and generating both text and images. Specifically, BLIP is trained on large-scale datasets containing paired text and image data, allowing it to learn the intricate relationships between visual and textual information. During the prompt generation process, BLIP uses its learned representations to generate coherent and contextually relevant text prompts based on the visual content of an image.

After generating the prompts, the model was trained for 2 epochs due to compute and time constraints with a dataset of 50K Imagenet images. The observed results indicated that the prompts were not well learned by the model, as most of the output images exhibited pixelation and lacked clear contour distinction. This suggested that more epochs may be required to effectively train the model with text prompts.

Additionally, an attempt was made to build a model that colorizes images based on the color scheme provided by an example image. This approach involved encoding the exemplar image using the CLIP model and incorporating it into the forward phase of the UNet. However, several constraints were encountered during the image encoding phase, preventing the completion of this approach.

Furthermore, an exploration into audio-to-image colorization was undertaken. The goal was to use audio mappings from videos to match the colorization scheme of images. To implement this, the pre-trained Wave2Vec2 model [BZMA20] was used to encode the audio input, followed by custom MLP layers to map the latent space of Wave2Vec2 and Stable Diffusion. Unfortunately, the model's performance was suboptimal, and effective learning did not occur.

Modality	FID Score	MSE
Unconditional	98.70	81.66
Text prompts	94.11	150.06

Table 1: FID Scores and MSE for unconditional (no text prompts) and text prompt based image colorization

Training Dataset size	FID Score	MSE
50,000	98.70	81.66
40,000	104.006	98.14
30,000	66.07	138.83

Table 2: FID Scores and MSE for unconditional (no text prompts) and text prompt based image colorization

4 Dataset

The ImageNet [DDS+09] dataset was fundamental to this research, providing the primary source of training data for the image colorization models. As a large-scale visual database designed for visual object recognition software research, ImageNet includes millions of annotated images across thousands of categories. A sample of 50,000 images from ImageNet was utilized to fine-tune the Control-Net model for unconditional image colorization. This extensive dataset enabled the model to learn detailed nuances in image structures and patterns, significantly enhancing its ability to colorize images accurately and vibrantly.
Additionally, for conditional colorization tasks, 50,000 images from ImageNet were used to train the model with text prompts, improving its responsiveness to textual inputs and overall performance.
A test set size of 64 images was used to compute metrics due to lack of computation resources.

5 Experiments

5.1 Unconditional vs Prompt-based Image Colorization

We compare the performance of unconditional image colorization with modality based image colorization by comparing FID scores produced by the models (15k samples for unconditional training, and 50,000 samples for conditional training). Images colorized without any text prompts and their corresponding ground truth images can be seen in Figure [3]. Images colorized with text prompts and their corresponding ground truth images can be seen in Figure [4]. We also tested with image sizes of 256 x 256 px, but they did not perform very well as compared to 512 x 512 sized images.

5.2 Studying the effect of dataset size on performance

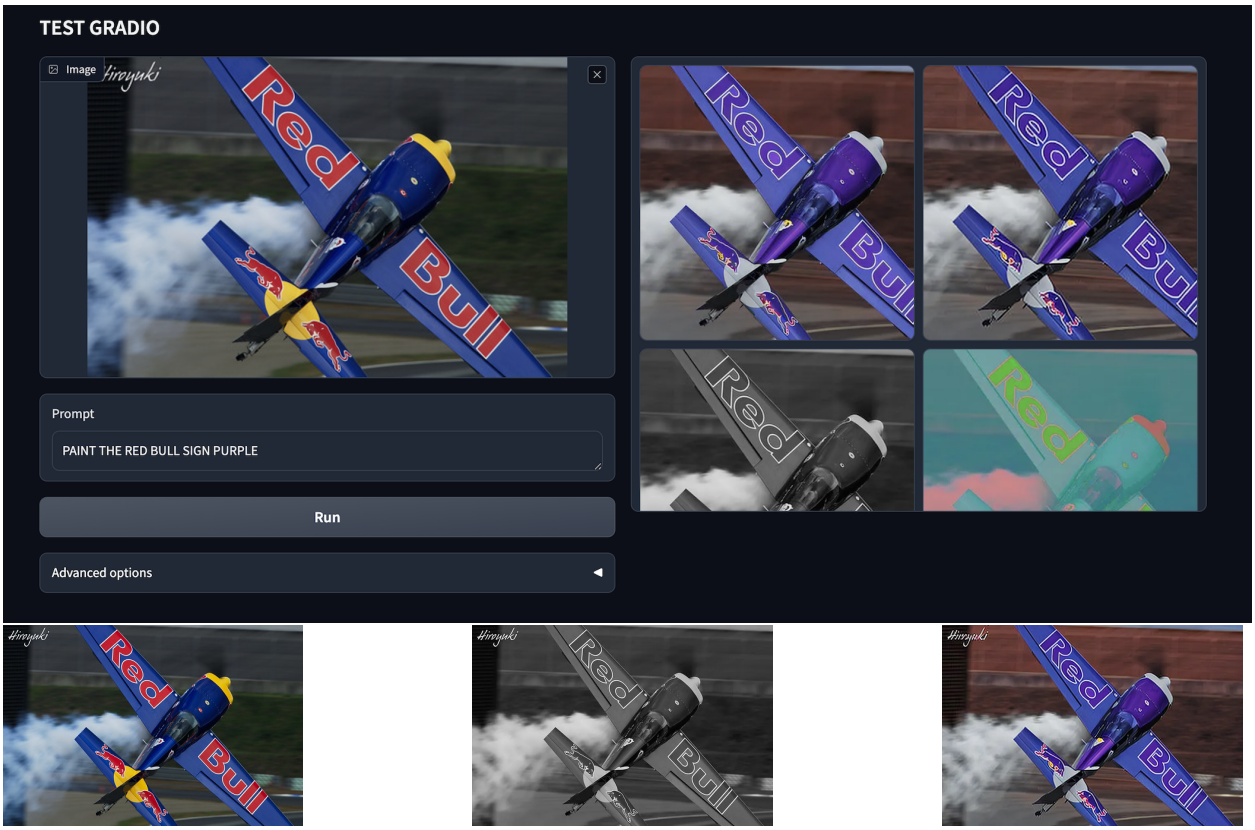
In order to understand the relationship between the dataset size and the quality of image colorization, we trained the models with dataset sizes of 50k, 40k and 30k samples respectively for 50,000 steps. The dataset samples in each split were shuffled, but every class was included in each split.

Fig. 3: Unconditional colorization (no text prompts). Top: Ground truth images. Bottom: Images colorized without text-prompts.

Fig. 4: Conditional colorization (using text prompts). Top: Ground truth images. Bottom: Images colorized with text prompts.

6 Results & Observations

Table 1 shows the FID scores and the Mean squared error (MSE) for a model trained using (a) unconditional data and (b) conditional data. The MSE loss increases for text-prompt based image colorization. Coupled with the image output samples, this implies that using text prompts needs more training time or a larger



(a) Input (b) Prepossessing (c) Colurized image

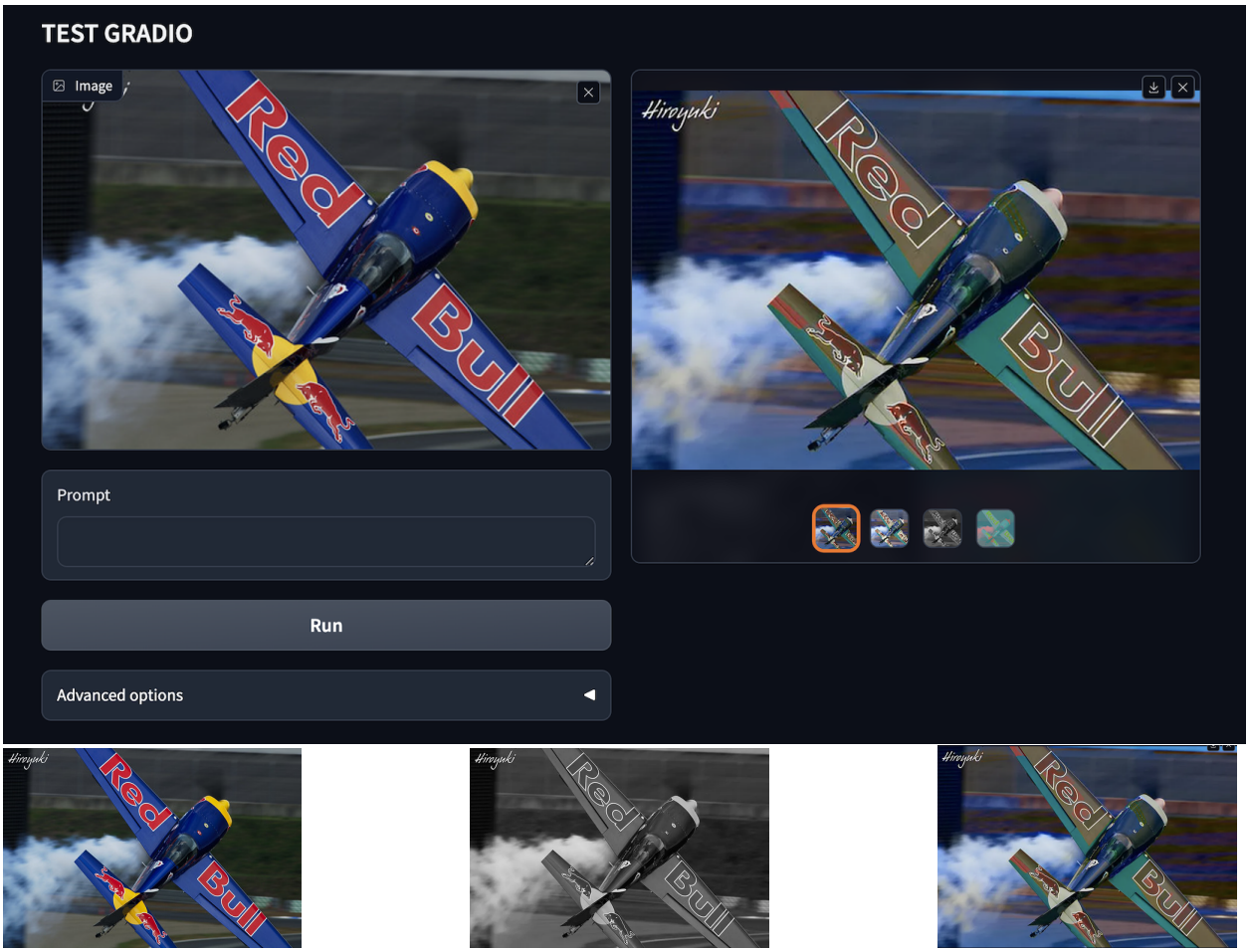
Fig. 5: Conditional on an image of an stuntplane

dataset. We were unfortunately unable to expand our experiments due to compute restraints. For experiments with no text prompts but varying training dataset size (Table 2), the MSE loss decreases with an increase in dataset size. This implies that given a fixed number of training steps, the model benefits with the increased dataset size.

The FID score shows no identifiable pattern in either of the experiments. [HRU+17] define the metric to be used on a sample size of 1024 images or more. Although modifications can be made to work on smaller dataset size, the resulting scores cannot be compared to other scores, and might also no longer correlate with visual quality. Given that we use a test set of 64 images due to resource constraints, we believe it would be best to only consider the FID metric when computed on a test set size of 1024 or more images.

Moving on to a more visual based evaluation approach, figures [5],[6],[7],[8] illustrates examples of conditional and unconditional image colourization where an input of airplane and cheetah are passed with the prompt (either filled for conditional or empty for unconditional). As a part of preprocessing, the image is first converted to its L-channel equivalent which is then passed to the trained model.

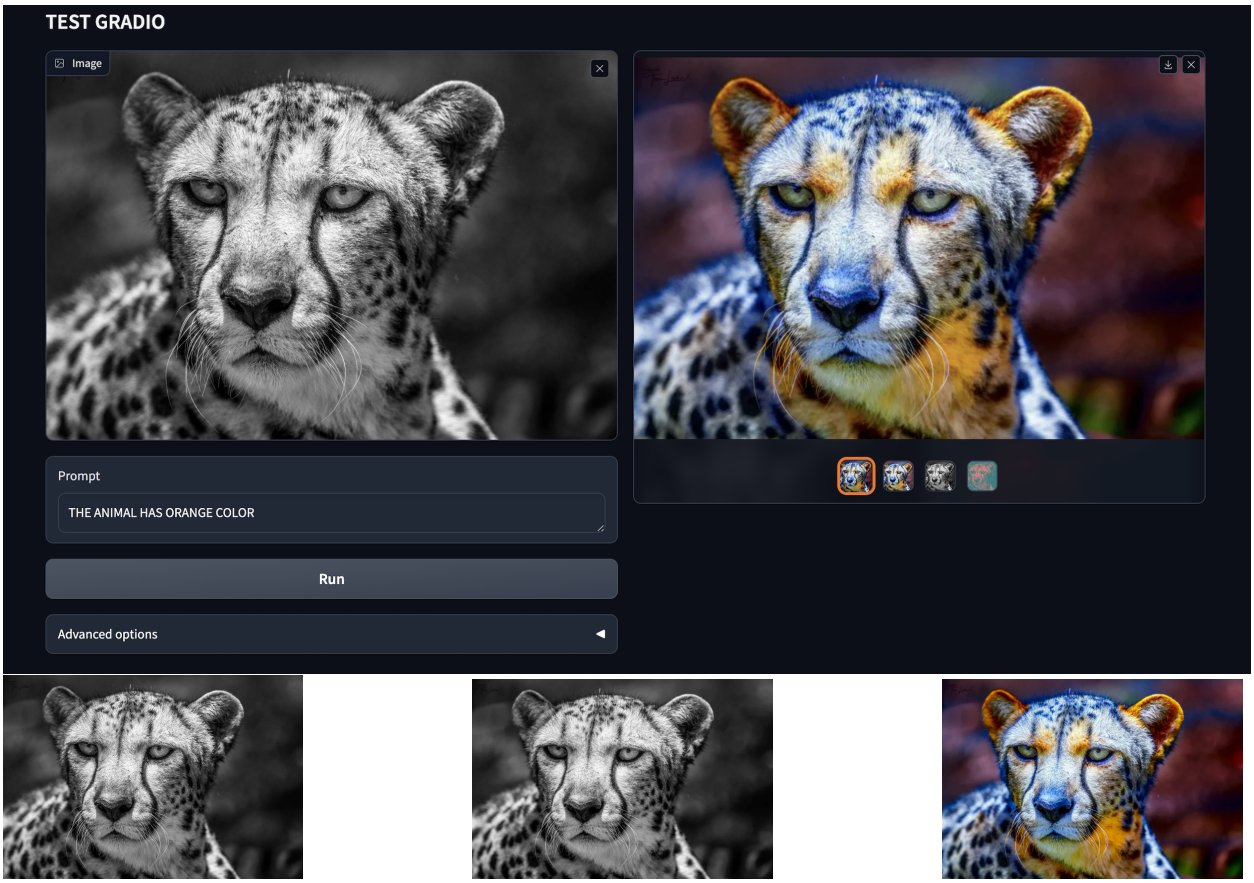
Throughout our project, we conducted inference using images of two different sizes: 512px and 256px.



(a) Input (b) Prepossessing (c) Colurized image

Fig. 6: Unconditional on an image of an stuntplane

Upon analyzing the results, we observed distinct differences in the quality of the colorization. Specifically, images with a resolution of 512px demonstrated significantly better segmentation, with sharp and well-defined contours, which contributed to a more accurate and visually appealing colorization. In contrast, images with a resolution of 256px exhibited noticeable color overflow, where the color distribution was less precise and often spilled over into adjacent areas, leading to a less refined and coherent visual output.(Refer figure [9] [10]).



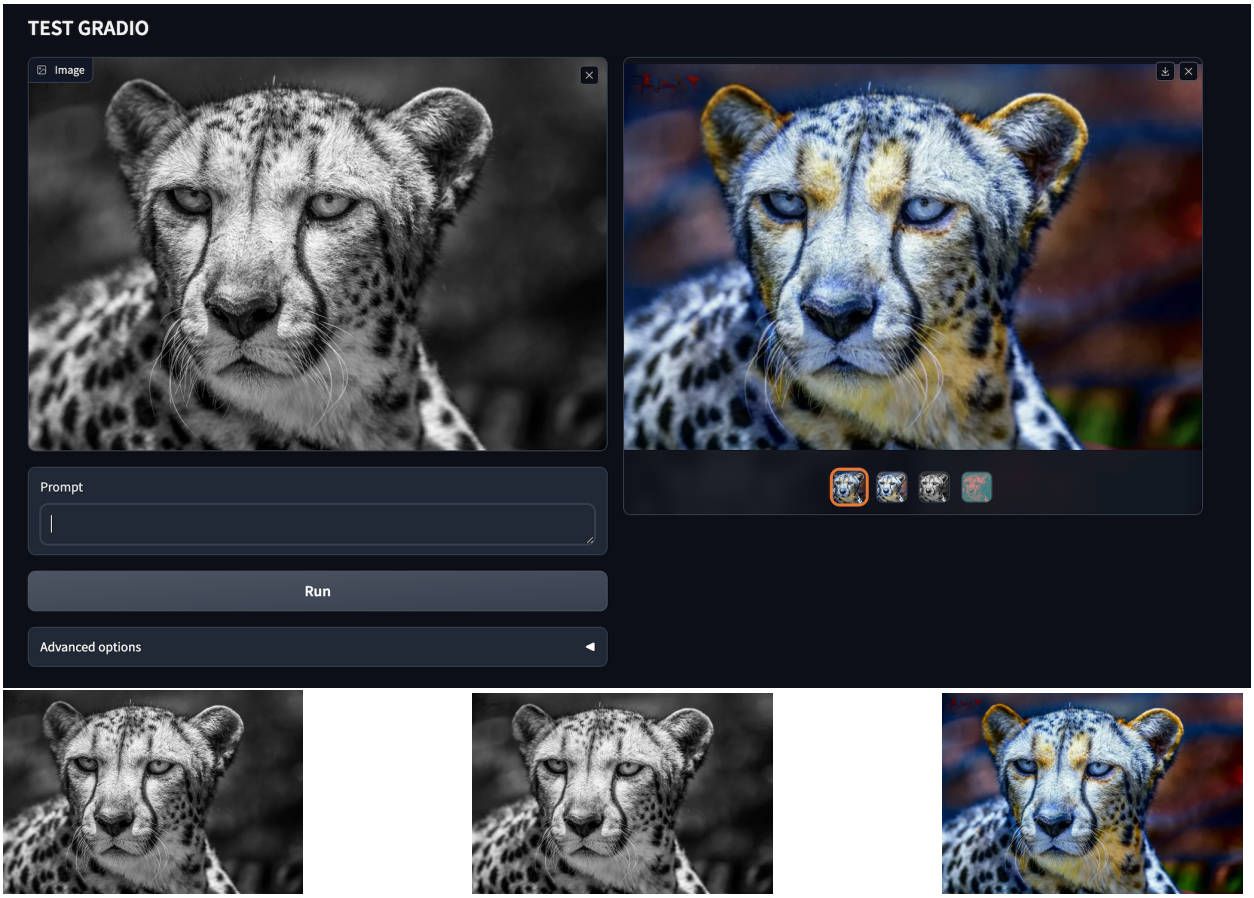
(a) Input (b) Prepossessing (c) Colurized image

Fig. 7: Conditional on a B/W image of a Cheetah

Conclusion & Future Work

By leveraging a substantial sample of 50,000 images from ImageNet, the Control-Net model was effectively fine-tuned for unconditional image colorization, demonstrating a remarkable ability to learn and replicate intricate image details and patterns. Experiments showed that images of size 512x512 outperformed those of size 256x256, displaying better segmentation and well-defined contours, while smaller images exhibited significant color overflow. These findings underscore the importance of high-quality, annotated data and appropriate image resolution in developing robust machine learning models.

Future work will focus on expanding the implementation of image colorization using audio prompts. This approach involves developing a separate model to convert audio files into their equivalent latent space representations, which can then be fed into the Control-Net model for colorization. Previous attempts encountered numerous bugs and dependency issues, but these can be addressed with additional resources.



(a) Input (b) Prepossessing (c) Colurized image

Fig. 8: Unconditional on a B/W image of a Cheetah

Acknowledgments

We would like to thank Prof. Saining Xie for teaching and giving an opportunity to work on this project topic. This experiment would not have been possible, if not for the NYU HPC and the team for supporting and resolving our issues. We would also like to thank the Imagenet Dataset and Stability AI for open sourcing stable diffusion model.

(a) Input (b) Conditional (c) Unconditional

Fig. 9: Image Inference with size 256 px

(a) Input (b) Conditional (c) Unconditional

Fig. 10: Image Inference with size 512 px

References

[BZMA20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.

Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[LLXH22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[LLZ+24] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization, 2024.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[ZAG+23] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion, 2023.

[ZIE16] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016.

[ZRA23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.