Expanding StyleCLIP: Enhancing Text-Driven Image Manipulation Beyond Facial Datasets
11-785 Introduction to Deep Learning Final Project Group 7

Pranav Setlur[1], Andy Tang[1], Youjeong Roh[2], and Minwoo Oh[2]

[1]Langauge Technology Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[2]Software and Societal Systems Department, School of Computer, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Deep learning models have revolutionized image generation and manipulation, with architectures like StyleGAN producing high-quality outputs. However, fine-grained semantic control remains a challenge due to entanglements within the model's latent space. StyleCLIP, introduced by Patashnik et al. in 2021, addresses this by combining the generative power of StyleGAN with the semantic understanding of CLIP (Contrastive Language Image Pre-training), enabling text-driven image manipulation. Despite its effectiveness, StyleCLIP's focus on identity preservation restricts its use to facial images, limiting its generalizability.

This project aims to overcome these limitations by introducing a modified loss function that prioritizes visual consistency over identity preservation. The model takes an image and a text prompt as inputs, optimizing the latent representation to produce an output image that aligns with the prompt while maintaining structural coherence. The model was evaluated on LSUN Horses and LSUN Churches, demonstrating improved adaptability and semantic precision for varied image types. This work broadens StyleCLIP's utility, paving the way for more flexible, text-guided image editing tools applicable to a wide range of real-world contexts.

1 Introduction

The ability to control and modify attributes of images generated by deep learning models has become a cornerstone in advancing AI-driven image editing. Generative Adversarial Networks (GANs) [3], particularly StyleGAN [7], allow for high-quality image generation, offering a foundation for tasks ranging from artistic creation to synthetic data generation. However, despite its strengths, StyleGAN struggles with achieving fine-grained, semantic-level control, as manipulations within its latent space often result in entangled modifications. This entanglement makes precise control difficult.

To address this challenge, StyleCLIP [11], introduced by Patashnik et al., integrates StyleGAN with Contrastive Language-Image Pre-training (CLIP) [12], facilitating intuitive, text-driven image manipulation within the latent space of StyleGAN. CLIP, trained on approximately 400 million image-text pairs from the internet, excels at understanding and encoding rich visual and semantic relationships. By combining the generative capabilities of StyleGAN with the semantic understanding of CLIP, StyleCLIP offers a powerful interface for modifying images based on natural language descriptions.

The primary input for StyleCLIP is an image and a text prompt describing the desired manipulation. The image is a latent vector in StyleGAN's W space. A latent vector is a mathematical encoding

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

of the image as a vector. The W space of StyleGAN is designed to provide more disentanglement compared to other latent spaces. This property makes it suitable for generating diverse and realistic outputs. The StyleCLIP model optimizes this latent vector to produce an output image that aligns with the given text prompt while maintaining the overall structure and identity of the original image. However, StyleCLIP primarily focuses on facial images, with its loss function heavily emphasizing identity preservation. This emphasis restricts its applicability to other image domains and creative use cases.

For this project, we address this limitation by enhancing StyleCLIP's versatility across various image domains. We introduced a novel loss function in the latent optimization process. This loss function was designed to prioritize visual consistency over identity preservation. This modification was implemented and tested to extend StyleCLIP's applicability to a wider range of datasets and image types beyond facial images. The datasets tested include architectural images (LSUN Churches) and animal images (LSUN Horses) [16].

Our work contributes to research in text-guided image manipulation and addresses a critical need for more flexible and generalizable models in computer vision and generative AI. The ability to manipulate diverse image types using natural language prompts represents a significant step towards more intuitive and accessible image editing tools.

2 Literature Review

2.1 Generative Adversarial Networks (GANs)

Generative models, particularly Generative Adversarial Networks (GANs) [3], have revolutionized the field of image synthesis since their introduction. GANs have evolved rapidly, with significant improvements in image quality and stability. StyleGAN [7] marked a significant advancement, gaining prominence for its ability to produce high-quality, diverse images by separating the latent space into distinct styles that control different aspects of the generated image. This allows for fine-grained manipulation. Subsequent iterations, StyleGAN2 [8] and StyleGAN3 [6], further improved image quality and addresses issues like aliasing. StyleGAN3, in particular, introduced translational equivariance, ensuring smoother interpolations and transformations. Despite these advancements, StyleGAN remains limited in achieving precise, semantic-level control, with manipulations often entangled. This limitation motivates the integration of complementary models, such as CLIP, to guide modifications using natural language prompts.

2.2 Language-Image Models

Contrastive Language-Image Pre-training (CLIP) [12] represents a significant advancement in visual and textual content understanding. CLIP is trained on a dataset of 400 million image-text pairs from the internet, enabling it to learn rich semantic relationships between visual and textual data. CLIP's architecture consists of two encoders, and they are trained to maximize the cosine similarity between the image and text embeddings of matching pairs while minimizing it for non-matching pairs. This contrastive learning approach allows CLIP to create a shared embedding space for both images and text.

The combination of StyleGAN and CLIP, as proposed in StyleCLIP [11], allows for more controlled image modifications by utilizing CLIP's embeddings to guide the manipulation of images within StyleGAN's latent space. This approach enables a more interactive and user-friendly interface to image editing, where users can provide text prompts to guide the manipulation process. This integration represents a significant step forward, addressing StyleGAN's limitations by introducing a natural language interface for attribute manipulation.

2.3 Loss Functions

StyleCLIP utilizes a combination of loss functions, including CLIP Loss for semantic alignment, Identity (ID) Loss to preserve facial identity, and L2 Regularization to maintain stability during optimization. While effective for facial images, this approach struggles to generalize to non-facial domains as the heavy emphasis on identity preservation becomes less relevant.

Alternative loss functions, such as perceptual loss [19] and GAN-based losses [8], have shown promise in improving the quality and accuracy of image manipulations. Perceptual loss captures high-level visual features, making it suitable for stylistic consistency, while GAN-based losses can enhance realism. Incorporating such losses could address StyleCLIP's current limitations and extend its applicability to diverse datasets.

2.4 Intersection of GANs and Language for Image Manipulation

Several works have explored the intersection of GANs and language for image manipulation. For example, DALL-E [13] generates images from text prompts. However, it cannot modify existing images directly. Similarly, GLIDE [9] emphasizes text-to-image generation but provides limited control for editing pre-existing content.

In contrast, StyleCLIP focuses on modifying existing images with precision, using text prompts for fine-grained control. This unique approach makes it valuable for high fidelity and semantic coherence applications, such as artistic editing or personalized content generation.

2.5 Recent Advances and Future Directions

Since the publication of StyleCLIP, researchers have actively sought to enhance its capabilities through various avenues. Key improvements include experimenting with alternative loss functions for better manipulation accuracy [15], integrating additional modalities such as depth or pose information [18], evaluating robustness across diverse datasets, including non-human images [1], and optimizing the framework for real-time applications [10].

Our project builds on these advancements by experimenting with different loss functions tailored to non-facial datasets. By introducing a novel loss function that prioritizes visual consistency, we aim to extend StyleCLIP's applicability to a broader range of domains.

3 Baseline Model Selection

Our project uses StyleCLIP [11] as the primary baseline model for image manipulation. StyleCLIP is a recent advancement in the field, combining the strengths of StyleGAN and CLIP to enable text-driven image editing capabilities. This model was introduced by Patashnik et al. in 2021 and has gained significant attention for its ability to enable intuitive manipulation of images based on natural language descriptions.

We selected StyleCLIP because of its innovative approach to integrating GANs with language-image understanding, making it a state-of-the-art solution for various image manipulation tasks. By leveraging CLIP's robust semantic embeddings alongside StyleGAN's high-quality image generation, StyleCLIP addresses limitations seen in previous models that lacked direct control over existing images. This capability is crucial for our goal of extending image manipulation techniques beyond facial datasets to encompass a wider range of image categories.

StyleGAN [7] generates images using a mapping network $f : Z \to W$ and a synthesis network $g : W \to X$, where Z is the initial latent space, W is the intermediate latent space, and X is the image space. The synthesis network uses adaptive instance normalization (AdaIN) for style control:

$$AdaIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i},$$

where x_i is the i-th feature map, and $y_{s,i}$ and $y_{b,i}$ are style and bias parameters.

CLIP [12] consists of an image encoder $E_I : X \to E$ and a text encoder $E_T : T \to E$, where T is the text space and E is the shared embedding space. CLIP is trained to maximize the cosine similarity between matched image-text pairs:

$$sim(i, t) = \frac{E_I(i) \cdot E_T(t)}{|E_I(i)||E_T(t)|}$$

StyleCLIP optimizes the latent code $w \in W$ to minimize the following loss function:

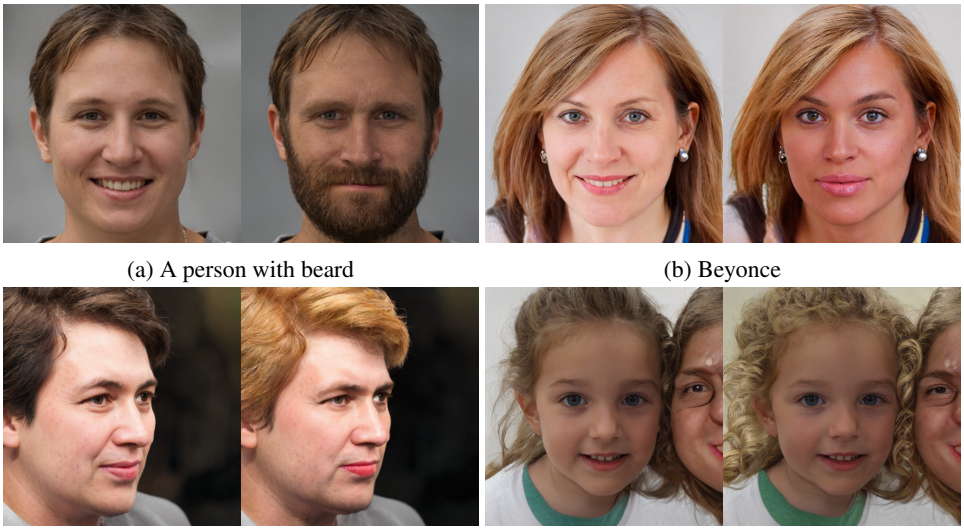
$$D_{\text{CLIP}}(G(w), t) + \lambda_{L2}\|w - w_s\|_2 + \lambda_{ID}L_{ID}(w),$$

where D_{CLIP} is the CLIP distance, w_s is the initial latent code, L_{ID} is an identity preservation loss, and λ_{L2} and λ_{ID} are weighting factors. The effectiveness of StyleCLIP is demonstrated through qualitative evaluations, particularly in user studies that assess the success of text-guided manipulations and the preservation of image quality.

4 Baseline Implementation and Analysis

To evaluate the baseline, we implemented StyleCLIP's latent optimization workflow using the original codebase provided by Patashnik et al. [11]. The model was applied to the FFHQ dataset [14], which consists of facial images. We confirmed the correctness of our implementation through qualitative evaluations, similar to the paper. Example outputs were visually inspected for semantic alignment with text prompts and preservation of image quality. The generated images demonstrate clear semantic transformations, such as "a person with a beard" or "Beyonce."

We observed convergence issues for specific text prompts when applied to different images during testing. To address this, we had to adjust the λ_{L2} parameter to emphasize semantic fidelity over latent stability. A visual comparison of the original and edited images is provided in Figure 1, showcasing successful text-guided manipulations across various domains.



(a) A person with beard (b) Beyonce

(c) Donald Trump (d) A person with curly hair

Figure 1: Baseline implementation with 4 prompts

5 Model Extensions

5.1 Methodology

We began by replicating the baseline latent optimization model outlined in the StyleCLIP paper. This served as the foundation for our proposed extensions. Figure 2 describes the process. The process begins with a latent vector w in StyleGAN's W space. This vector encodes the input image's features. StyleGAN generates an output $G(w)$ based on this latent vector. To align the output with a desired textual modification ("without makeup" in the figure), the model uses the CLIP embedding space. CLIP consists of two encoders: the text encoder, which maps the input text prompt to a semantic embedding, and the image encoder, which maps the generated image to a corresponding embedding. The similarity between these

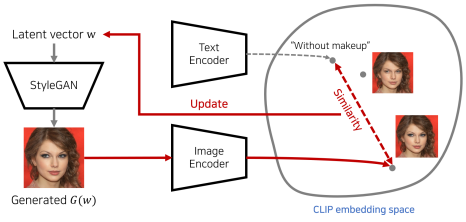


Figure 2: Latent Optimization Process

embeddings is the primary optimization criterion. During optimization, w is iteratively updated to minimize the CLIP similarity distance between the generated image embedding and the target text embedding.

To extend StyleCLIP's applicability to non-facial datasets, we proposed a modified loss function to improve generalization. The original loss function optimizes the alignment between the generated image and target text prompt while maintaining identity and visual similarity. It is given by:

$$\arg\min_{w \in W} D_{\text{CLIP}}(G(w), t) + \lambda_{L2}\|w - w_s\|_2 + \lambda_{ID} L_{ID}(w),$$

where $w, w_s \in W$ are the modified and original image in StyleGAN's latent space, $D_{CLIP}(G(w), t)$ measures the cosine distance between CLIP embeddings of the generated image and text prompt t, and $L_{ID}(w)$ is an identity loss generated by an ArcFace model.

Our modification replaced the identity loss $L_{ID}(w)$ with a generalized image consistency loss:

$$\arg\min_{w \in W} D_{\text{CLIP}}(G(w), t) + \lambda_{L2}\|w - w_s\|_2 + \lambda_{IM} D_{\text{CLIP}}(G(w), G(w_s)),$$

where $D_{\text{CLIP}}(G(w), G(w_s))$ measures the cosine distance between CLIP embeddings of the generated and original images. This modification reduced the emphasis on facial identity preservation, allowing the model to maintain structural consistency across diverse datasets.

5.2 Datasets

Dataset	Type	Resolution	Size
FFHQ	Facial Images	1024×1024	$70,000$
CelebA-HQ	Celebrity Faces	1024×1024	$30,000$
LSUN Horses	Animal Images	256×256	$120,000+$
LSUN Churches	Architectural Images	256×256	$160,000+$

Table 1: Datasets tested

In the StyleCLIP paper, StyleGAN models are pre-trained on Flickr-Faces-HQ (FFHQ) [14], a dataset consisting of 70,000 high-resolution facial images with a 1024×1024 resolution. FFHQ provides a diverse set of facial attributes such as age, ethnicity, and accessories (e.g., glasses, hats), making it highly suitable for training models on a range of facial manipulation tasks.

For image manipulations, StyleCLIP uses CelebA-HQ [5], a high-quality version of the CelebA dataset, featuring 30,000 celebrity face images at a 1024×1024 resolution. Each image in CelebA-HQ is annotated with 40 attributes, including gender, smile, and facial hair, which facilitates training models to target specific facial features while retaining photorealism. Images are inverted using Encoder for Editing (e4e) to bring them back into the latent space of StyleGAN for more effective manipulation.

To evaluate our modified model, we extended the scope of StyleCLIP beyond facial image datasets by incorporating LSUN Horses and LSUN Churches [16]. These datasets were selected to challenge the model's capability in non-human, non-facial image manipulation tasks.

- **LSUN Horses:** Contains high-quality images of horses in diverse natural settings. This dataset evaluates the model's ability to perform text-guided manipulations while preserving complex textures and backgrounds.
- **LSUN Churches:** Includes a wide variety of architectural styles and lighting conditions. This dataset enables the assessment of StyleCLIP's performance on structured, non-living objects.

The datasets were preprocessed to fit the input requirements of StyleGAN models. This involved resizing images to a 1024×1024 resolution and normalizing pixel intensities. Images were inverted into the latent space of StyleGAN using e4e to facilitate efficient manipulations in the latent domain. For instance sampling, we employed uniform sampling across the datasets to ensure diverse text

prompts and visual compositions during training and evaluation. This approach also mitigated potential biases that might arise from over representation of specific styles or textures in the dataset.

These datasets were selected to challenge StyleCLIP's capacity for diverse, non-facial image manipulation. These results confirmed the robustness and versatility of our modified model, particularly in scenarios that extend beyond the original focus on human faces.

5.3 Evaluation Metrics

We employed qualitative evaluation metrics similar to those used in the original StyleCLIP paper [11]. These included visual comparisons and user studies to assess:

- **Semantic Accuracy:** How well the generated image aligns with the text prompt.
- **Image Quality:** Realism and coherence of the manipulated images.
- **Preservation of Key Attributes:** Retention of structural consistency across diverse datasets.

While quantitative metrics such as Frechet Inception Distance (FID) could be explored in future work, our primary focus was on qualitative assessments, similar to Patashnik et al., due to their relevance in evaluating semantic alignment and manipulation quality.

6 Experiments and Results

Our experiments evaluated the performance of the modified StyleCLIP model on both the original FFHQ datasets and non-facial datasets (LSUN Horses and Churches). The primary goals were to demonstrate improvements in image quality, structural consistency, and adherence to text prompts, particularly for non-facial datasets where the baseline StyleCLIP struggles.

6.1 Results on FFHQ Dataset

The FFHQ dataset provides a diverse set of high-resolution facial images. Figure 3 compares the original StyleCLIP method with our proposed method using prompts such as "A person with blond hair" and "A person with green hair."

While the baseline approach performs adequately, it often produces blurry outputs and struggles with fine-grained alignments to text prompts. For instance, in Figure 3c, the baseline method produces inconsistent coloring and does not manipulate the image based on the prompt. On the other hand, our approach generates sharper images with better adherence to the prompts while preserving the original structure and texture of the images.

We empirically found that proposed approach is more powerful on creative prompts. This is because our loss function focuses more on visual similarity than identity preservation. StyleCLIP fails to generate coherent images when given creative prompts like "A person with beard" to females. However, for prompts including common features, like "blond hair," StyleCLIP does better because it preserves the facial identity well. Figure 3 illustrates this.

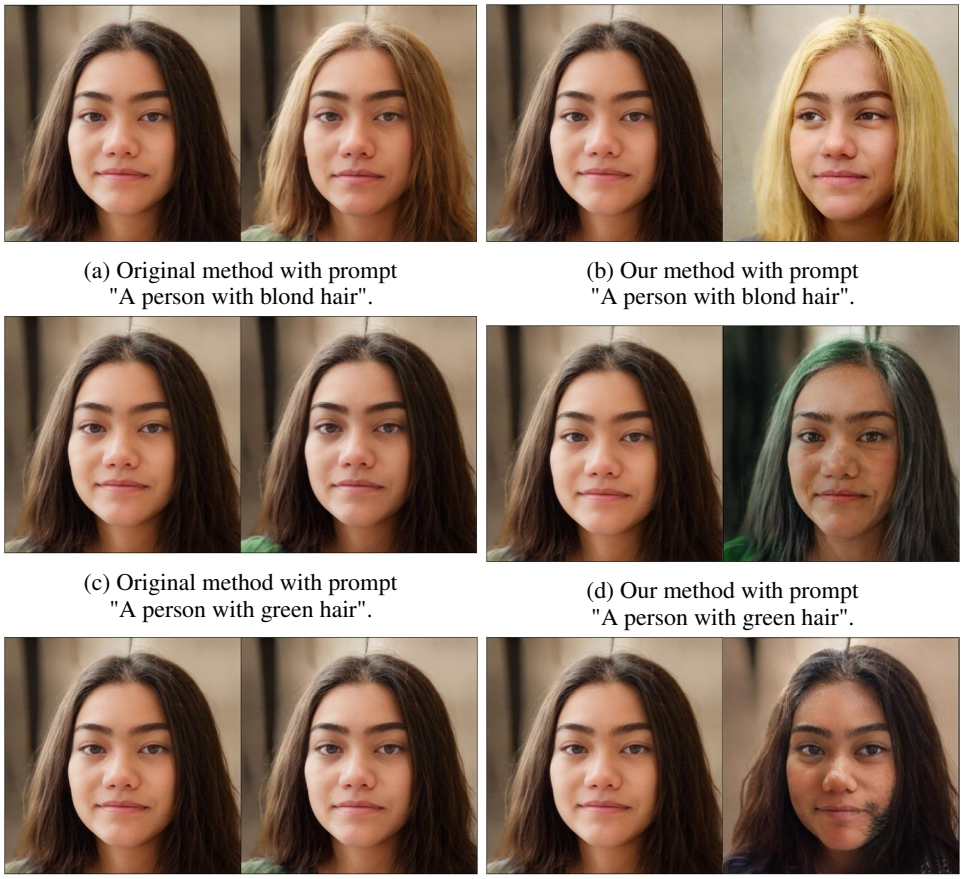
6.2 LSUN Horses

The LSUN Horses dataset was chosen to test the model's performance on images featuring non-human subjects with complex textures and backgrounds. Figure 4 shows results for prompts such as "golden horse" and "watercolor painting."

The baseline StyleCLIP model struggles with these prompts, often distorting textures or failing to preserve background details. For example, in figure 4a, our method applies the color change more naturally, preserving the horse's structure better than the original method. Our method demonstrates a clear advantage by producing visually consistent outputs. The generated images retain the rich textures and natural backgrounds while adhering more closely to the prompts.

6.3 LSUN Churches

To assess performance on structured architectural datasets, we applied our model to LSUN Churches. Figure 5 compares results for prompts such as "starry night" and "igloo."



(a) Original method with prompt "A person with blond hair".

(b) Our method with prompt "A person with blond hair".

(c) Original method with prompt "A person with green hair".

(d) Our method with prompt "A person with green hair".

(e) Original method with prompt "A person with beard".

(f) Our method with prompt "A person with beard".

Figure 3: Result comparison on FFHQ facial dataset

The baseline model again exhibits limitations, with outputs that are either over-smoothed or contain visual artifacts. For instance, the prompt in figure 5e leads to poorly defined structures in the baseline output, where integrating the prompt's artistic style into the church architecture is inconsistent.

Our method successfully captures the structural details of the churches while incorporating stylistic features from the prompts. This demonstrates the robustness of our generalized consistency loss in handling complex, non-facial datasets.

6.4 Qualitative Metrics

To further substantiate our results, we qualitatively assessed the fidelity of the generated images based on:

- **Adherence to Prompts:** The extent to which the generated images align with the given text descriptions.
- **Visual Coherence:** The degree to which outputs maintain natural textures, structural integrity, and photorealism.
- **Comparative Evaluation:** Side-by-side comparisons of outputs generated by the baseline and proposed methods.

Table 2 summarizes qualitative improvements across datasets, highlighting the advantages of our approach.

(a) Original method with prompt "golden horse".

(b) Our method with prompt "golden horse".

(c) Original method with prompt "watercolor painting".

(d) Our method with prompt "watercolor painting".

(e) Original method with prompt "alpaca".

(f) Our method with prompt "alpaca".

Figure 4: Result comparison on LSUN Horse dataset

Dataset	Baseline Fidelity	Proposed Fidelity	Baseline Coherence	Proposed Coherence
FFHQ	Moderate	High	High	High
LSUN Horses	Moderate	High	Low	High
LSUN Churches	Low	High	Low	High

Table 2: Qualitative Metrics

7 Discussion

Our project aimed to extend StyleCLIP's capabilities by modifying its loss function and evaluating its performance on diverse datasets beyond facial images. The results of our experiments uncovered both promising capabilities and areas requiring refinement.

7.1 Relevance of Findings

Our experiments revealed that the modified loss function enhances StyleCLIP's applicability to non-facial datasets, outperforming the original model in preserving semantic alignment while maintaining visual consistency. The inclusion of a generalized visual similarity term ($\lambda_{IM} D_{\text{CLIP}}(G(w), G(w_s))$) allowed the model to handle stylistically varied datasets like LSUN Horses and LSUN Churches. The results validate that reducing emphasis on identity preservation broadens the model's utility for tasks beyond facial image manipulation. This makes the model a versatile tool for creative industries and domain-specific applications.

(a) Original method with prompt "starry night".

(b) Our method with prompt "starry night".

(c) Original method with prompt "igloo".

(d) Our method with prompt "igloo".

(e) Original method with prompt "wood cabin".

(f) Our method with prompt "wood cabin".

Figure 5: Result comparison on LSUN Church dataset

Our ablation study further revealed the critical role of the λ_{L2} term in the loss function. As illustrated in figure 8, varying λ_{L2} significantly impacts the model's performance. A high value overly penalizes deviations, resulting in outputs that lack meaningful stylistic changes. Conversely, setting it too low leads to outputs that deviate excessively from the input, undermining coherence. This indicates that careful tuning of λ_{L2} is necessary to effectively balance visual consistency and stylistic transformation.

7.2 Sensitivity of Results

The results were susceptible to certain hyperparameters and input variations:

- **Lambda values in Loss Function:** λ_{L2} values require precise tuning. For instance, experiments showed that $\lambda_{L2} = 6 \times 10^{-4}$ on LSUN horses yielded little to no change while $\lambda_{L2} = 5 \times 10^{-5}$ resulted in incoherent, overly stylized outputs (see Figure 8). We found that $\lambda_{L2} = 2 \times 10^{-4}$ worked the best for LSUN horses and churches.

  Similarly, $\lambda_{IM} = 1$ or 2 yielded little to no change, while $\lambda_{IM} = 0$, as in the original method, yielded visually inconsistent results (see Figure 9). We found $\lambda_{IM} = 0.5$ to perform the best across all datasets.

- **Learning Rates:** We used a cosine decay learning rate schedule across all our experiments. On LSUN horses and LSUN churches, too low an initial learning rate (e.g., 0.1) did not result in any meaningful change to the images, while too high a learning rate (e.g., 0.8) led to poor convergence. We found an initial LR of 0.4 to perform the best in our experiments. We believe that a higher learning rate perturbs the latent vector more, which allows images

Figure 6: Effect of λ_{L2} on results.
Sequentially, λ_{L2} = 0.08, 0.008, 0.0008.
The pictures above are the original method, the pictures below are our method. The prompt was "A person with blond hair".

to be pushed towards more creative prompts. A simpler prompt such as "golden horse" is able to be trained with a lower learning rate, while "alpaca" requires a higher learning rate.

- **Prompt-Image Interactions:** Some prompts yielded better results for specific images. This suggests a dependence on the inherent style and features of the input image. For instance, "golden horse" worked better on brighter images with simple backgrounds, while more complex prompts struggled with detailed scenes. Creative prompts such as "igloo" simply did not work on some input images.

7.3 Risks and Uncertainties

Despite promising results, certain risks and uncertainties must be addressed:

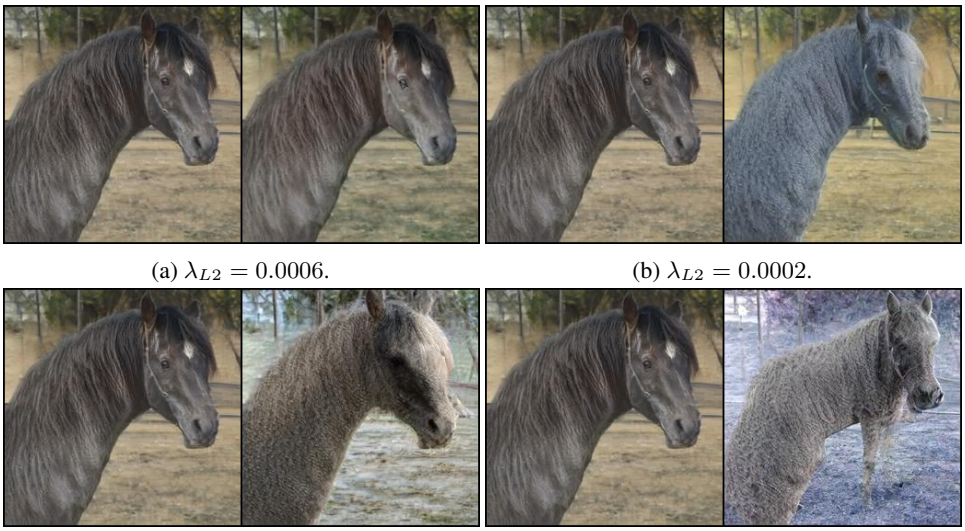
- **Dataset Bias**: The diversity and representation of our training datasets may introduce biases in the model's performance across different images and styles.
- **Loss Function Trade-Offs**: The shift from identity preservation to generalized visual similarity introduces potential risks in applications requiring fidelity to original subject characteristics, such as personalized content generation.
- **Parameter Sensitivity**: The strong dependence on hyperparameter tuning, such as λ_{L2} and learning rate, introduces variability in performance, especially for previously unseen domains or styles.

7.4 Comparative Analysis

Our experiments demonstrate that our modified StyleCLIP model generally outperforms the original in non-facial domains, particularly for complex textures and architectural images. However, the performance varies depending on the specific prompt and image type.

Our model shows comparable or slightly improved performance for facial images, particularly in maintaining overall image quality while applying text-guided manipulations. Our model consistently produces more coherent and visually appealing results in non-facial domains, especially for style transfer tasks (e.g., "igloo" prompt).

Figure 7: Effect of λ_{IM} on results.
Sequentially λ_{IM} = 0.0001, 0.001, 0.002.
The pictures above are the original method, the pictures below are our method. The prompt was "A person with blond hair." The original method is unaffected byλ_{IM}.



(a) $\lambda_{L2} = 0.0006$.

(b) $\lambda_{L2} = 0.0002$.

(c) $\lambda_{L2} = 0.0001$.

(d) $\lambda_{L2} = 0.00005$.

Figure 8: Effect of λ_{L2} on results

However, on facial datasets, the original StyleCLIP occasionally produced more natural-looking outputs for identity-focused tasks, such as hair color changes. This could be because the fidelity to input features is crucial. This highlights a trade-off between generalization and identity preservation.

8 Future Work

Our current work presents a strong foundation for extending StyleCLIP to non-facial datasets, but future efforts should address a few limitations and areas for improvement. These include enhancing

(a) $\lambda_{IM} = 0$.

(b) $\lambda_{IM} = 0.5$.

(c) $\lambda_{IM} = 1$.

(d) $\lambda_{IM} = 2$.

Figure 9: Effect of λ_{IM} on results

model performance, increasing robustness across datasets, and refining experimental methodologies to achieve state-of-the-art results.

Despite promising results, our model does not consistently outperform StyleCLIP in all scenarios. For example, in experiments with facial datasets, our method exhibited performance comparable to StyleCLIP but struggled with certain prompts. The results suggest a need for further optimization in tasks that require fine-grained feature manipulation. We have identified a few areas for improvement.

8.1 Exploring Alternative CLIP Models

While the original CLIP embeddings formed the foundation of our approach, future experiments should investigate alternative embedding models such as OpenCLIP [4] or Florence [17]. These models are trained on broader datasets and employ architectural advancements that may yield higher-quality embeddings for complex prompts. For instance, Florence's multimodal capabilities could better capture nuanced relationships between text and image styles, potentially reducing alignment errors.

8.2 Expanding Dataset Diversity

A significant limitation of our study was the relatively narrow dataset selection. While LSUN Horses, Churches, and FFHQ provided valuable benchmarks, they do not encompass the full range of artistic, natural, and synthetic image domains. Expanding the dataset to include abstract art, scientific visualizations, and culturally significant artifacts could provide a more comprehensive evaluation. Training on such datasets can broaden the model's applicability to specialized and real-world tasks. In turn, this will enhance its robustness and reduce dataset bias.

8.3 Refining Loss Functions

Our reliance on a generalized consistency loss proved beneficial but introduced limitations in identity preservation tasks. Future work should explore hybrid loss functions that incorporate perceptual loss [19] for improved human-aligned preferences, GAN losses [8] for enhanced visual fidelity, or attribute-specific losses tailored to critical features such as texture consistency or fine-grained attributes. These modifications could reduce artifacts and improve outputs in both artistic and structural domains.

8.4 Evaluation Metrics for Diverse Applications

Our current evaluation relies heavily on qualitative assessments. Future work should incorporate a more comprehensive set of quantitative metrics. This could include perceptual metrics like R-LPIPS [2] and domain-specific metrics for different types of images. For instance, metrics that quantify structural preservation could be developed when manipulating architectural images. As for prompt adherence, one possible method is to develop a golden edited image through some means, from which images can be compared to. A crafted dataset of this kind would be immensely helpful in evaluating the quality of edited images.

Future research can enhance the capabilities and adaptability of StyleCLIP modifications by addressing these limitations and exploring new avenues. These improvements will push toward state-of-the-art performance and unlock applications across creative, scientific, and industrial domains.

9 Conclusion

This project aimed to address StyleCLIP's limitations, particularly its focus on facial images and reliance on identity preservation, which restricted its applicability to diverse image types. Our goal was to extend StyleCLIP's capabilities to a broader range of visual domains while maintaining high-quality, text-guided image manipulations.

To achieve this, we modified StyleCLIP's loss function to prioritize visual consistency over identity preservation. This modification enhanced the model's applicability to broader image datasets beyond faces. Our experiments demonstrated significant improvements in the model's flexibility. We were able to generate high-quality, coherent images across diverse domains such as architectural structures (LSUN Churches) and animals (LSUN Horses), in addition to human faces (FFHQ).

The results of our study have important implications for AI-driven image editing. By successfully extending StyleCLIP's capabilities to non-facial datasets, we have expanded the potential applications of text-guided image manipulation. This advancement opens up new possibilities in various fields, including digital art, content creation, and visual effects.

While our modified model showed clear improvements over the original StyleCLIP in many scenarios, particularly with non-facial images, we acknowledge that there is still room for further enhancement. Our work achieved the primary goal of extending StyleCLIP's applicability, but we identified areas for improvement, such as consistency across different prompts and fine-grained control in certain scenarios.

This project represents a significant step forward in making text-guided image manipulation more versatile and accessible across a broader range of visual domains. It lays the groundwork for future research to further refine and expand these capabilities, potentially leading to more powerful and flexible tools for creative and practical image manipulation applications.

Division of Work

Pranav worked on modifying the loss function, writing large portions of the report, and creating the project video. Andy worked on generating images for the LSUN horse and LSUN church datasets as well as tuning hyperparameters. Youjeong worked on inverting datasets with e4e as well as generating images for the FFHQ datset. Minwoo helped with image generation and writing the report.

GitHub Repository

References

[1] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. Enhancing multimodal understanding with clip-based image-to-text transformation. In *Proceedings of the 2023 6th International Conference on Big Data Technologies*, ICBDT '23, page 414–418, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, and Alexandre Araujo. R-lpips: An adversarially robust perceptual similarity metric. *ArXiv*, abs/2307.15157, 2023.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[4] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[6] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc.

[7] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228, December 2021.

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021.

[10] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2065–2074, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[14] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), Oct. 2016.

[15] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, 2021.

[16] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.

[17] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021.

[18] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6868–6874, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.