Text Guided Image Modification

Group Members

Atharv Kumar (B21038)
Krishna Kumar Dixit (B21013)
Mourya Kondawar(B21016)
Yashika Gupta (B21174)

Faculty Guide: Prof. Parimala Kancharla

Introduction

- Text-conditioned image editing has recently attracted considerable interest. However, most methods are currently limited to one of the following: specific editing types (e.g., object overlay, style transfer), synthetically generated images, or requiring multiple input images of a common object. Unlike prior methods that either work with synthetic images, require multiple inputs, or support limited types of edits, Imagic allows non-rigid, text-based editing of a single real image. It preserves the core characteristics of the original image while enabling transformations such as changing posture, composition, or adding elements.

- We aim to demonstrate complex text-based semantic edits to a single real image. For example, we can change the posture and composition of one or multiple objects inside an image, while preserving its original characteristics.

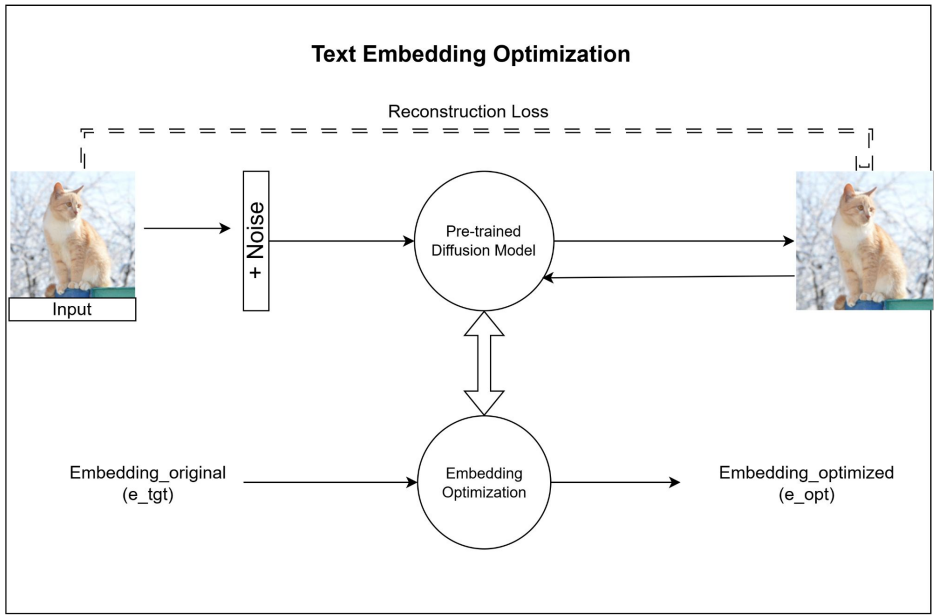
Objectives

- **Main Goal:**
  To develop a versatile and efficient method for editing real images using natural language descriptions.

- Enable complex, non-rigid edits on a single high-resolution real image eg change object posture or composition, add, remove, or modify objects.
- Achieve high-quality edits that align closely with the target text description, preserve the original image's details, such as background and object identity.
- Overcome limitations of existing methods like there should be no need for additional inputs like masks or multiple images.
- Leverage advanced AI diffusion models to ensure realism in the output.

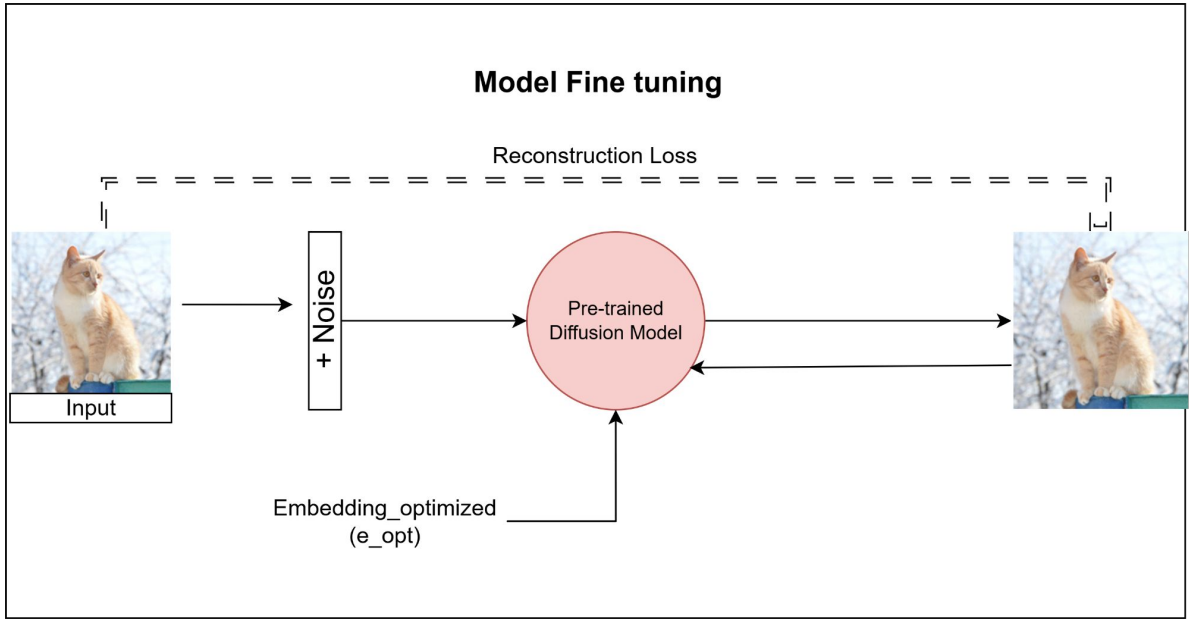
Model Flowchart

PART A : TEXT EMBEDDING OPTIMIZATION



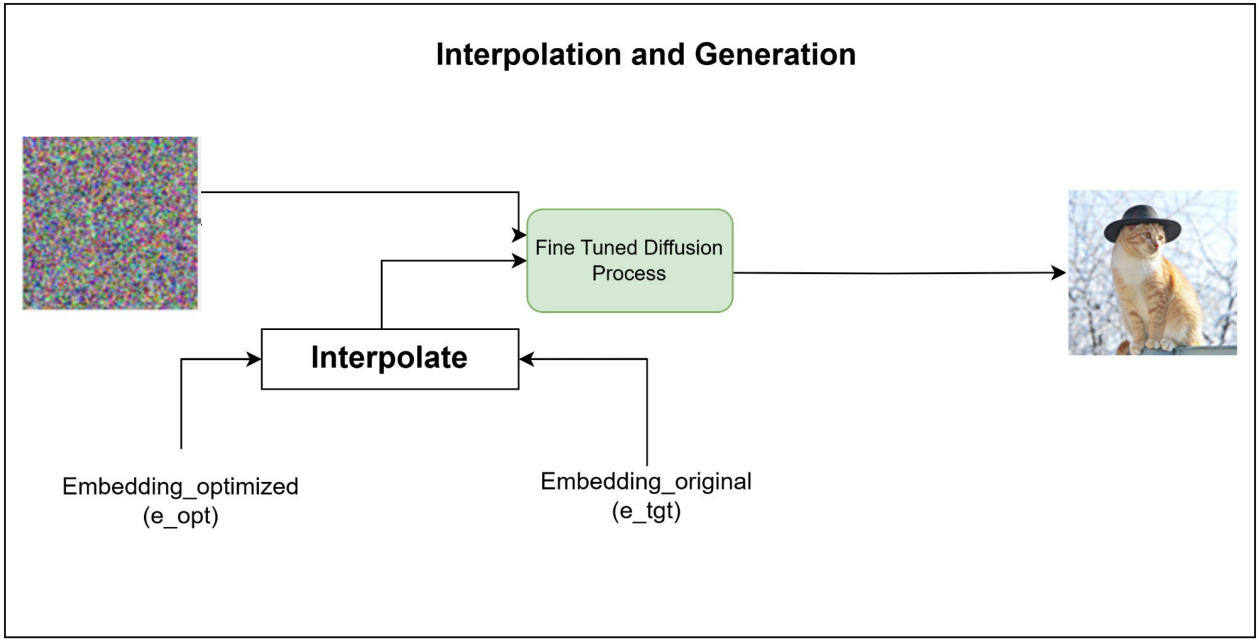
Model Flowchart

PART B : MODEL FINE TUNING



Model Flowchart

PART C : INTERPOLATION AND GENERATION



Existing Methods

GAN-Based Methods:

- Utilize Generative Adversarial Networks for edits.
- Struggle with real image fidelity and complex transformations.

Latent Space Manipulation:

- Adjust features within pre-trained model latent spaces.
- Often limited to simple edits and require extensive optimization.

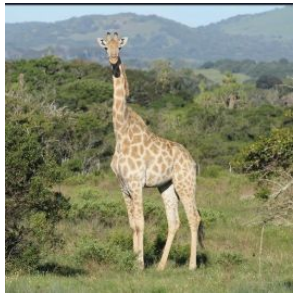
Diffusion-Based Models:

- Use iterative noise-based processes to edit images.
- Examples:
    - *SDEdit*: Focuses on global edits with added noise but lacks fine-grained control.
    - *DDIB*: Edits images through inversion but needs auxiliary inputs.
    - *Text2LIVE*: Good for localized edits but cannot handle complex non-rigid transformations.

Dataset/Samples

- We use a dataset named **TEdBench** to evaluate the model.

1. **Input_list.json** : Original Image name, target_text
2. **originals :** Directory of original images




```
{
  "img_name": "apples.jpeg",
  "target_text": "A basket of oranges."
},
```

```
{
  "img_name": "apples.jpeg",
  "target_text": "A bowl of apples."
},
```

```
{
  "img_name": "giraffe.jpeg",
  "target_text": "A giraffe with a short neck."
},
```

```
{
  "img_name": "giraffe.jpeg",
  "target_text": "A photo of a giraffe eating the grass below."
},
```

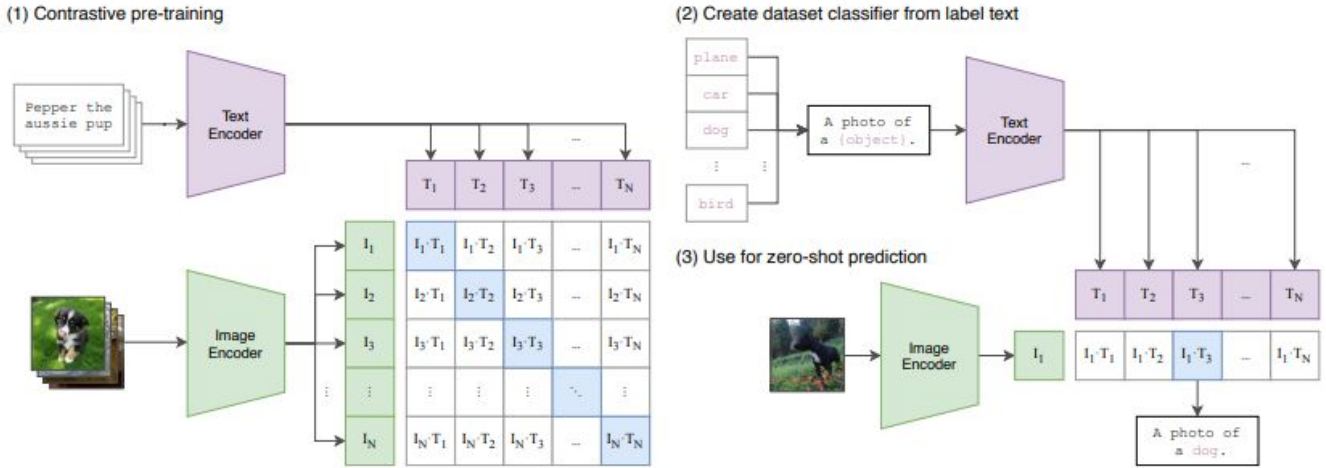
```
{
  "img_name": "dog2_standing.png",
  "target_text": "A photo of a sitting dog."
},
```

```
{
  "img_name": "dog2_standing.png",
  "target_text": "A photo of a jumping dog."
},
```

Methodology

Text Embedding Optimization:

As a first step we plan to get target_text embeddings (e_tgt) and later align the target text with the input image to get meaningful insights (e_opt). We have used CLIP(openai/clip-vit-large-patch14) model for the same. **[Average Cosine Similarity : 76.45%]**

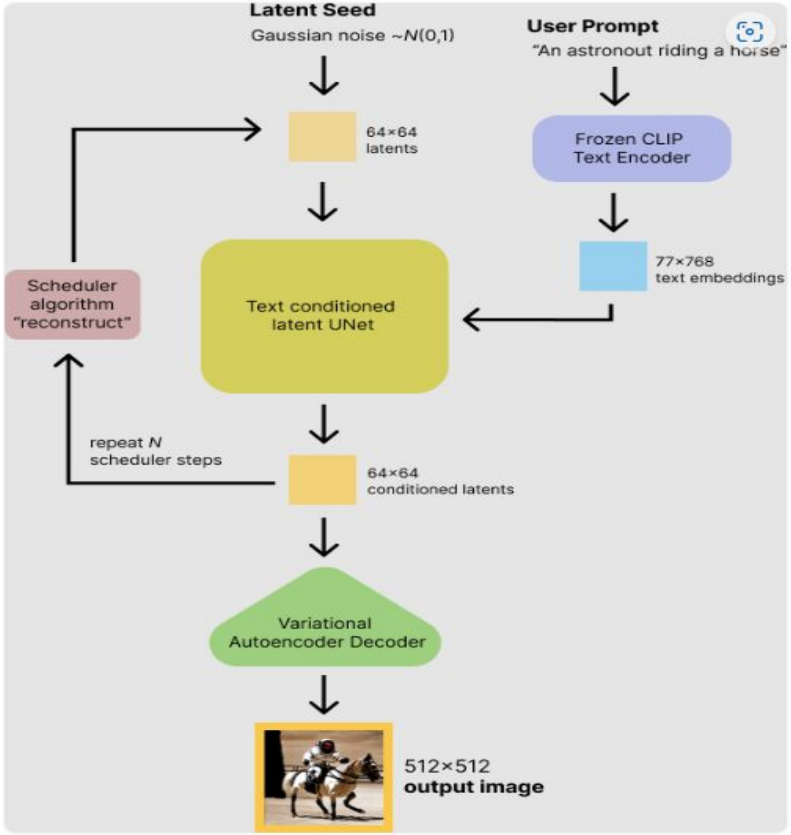


Methodology

Model Fine Tuning

This process shifts the model to fit the input image x at the point e_opt. In parallel, we fine-tune any auxiliary diffusion models present in the underlying generative method, such as super-resolution models. We fine-tune them with the same reconstruction loss, but conditioned on e_tgt, as they will operate on an edited image. This is done to ensure that model retains details of the input image during edits.

MSE(Mean Squared Error) [Reconstruction Loss]: 5.81%



Methodology

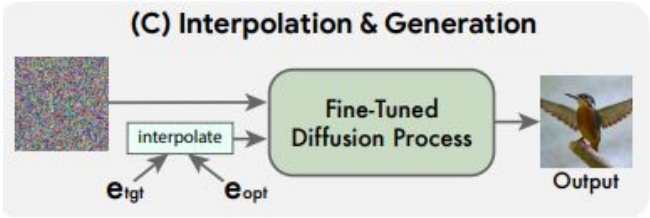
Interpolation and Generation

Since the generative diffusion model was trained to fully recreate the input image x at the optimized embedding e_opt, we use it to apply the desired edit by advancing in the direction of the target text embedding e_tgt. More formally, our third stage is a simple linear interpolation between etgt and e_opt.

For a given hyperparameter η:

$$\bar{\mathbf{e}} = \eta \cdot \mathbf{e}_{tgt} + (1 - \eta) \cdot \mathbf{e}_{opt},$$

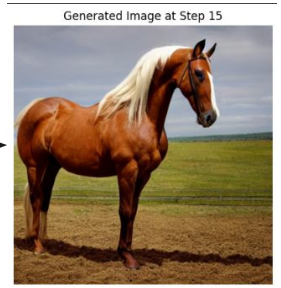
We then apply the base generative diffusion process using the fine-tuned model, conditioned on $\bar{e}$. This results in our edited image



Results

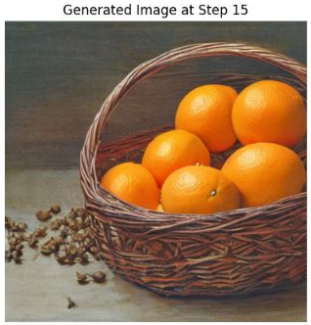
Evaluation Metric : CLIP Score (1 : Perfect Alignment of Image and Text , 0 : No Alignment , -1 : Opposite Alignment)
SSIM Score (Alignment with the Input Image and the generated Image)

(target_text)
A horse standing



CLIP Score : 0.3951
SSIM Score : 0.2457

(target_text)
A basket of oranges



CLIP Score : 0.4137
SSIM Score : 0.1014

Results

Evaluation Metric : CLIP Score (1 : Perfect Alignment of Image and Text , 0 : No Alignment , -1 : Opposite Alignment)
SSIM Score (Alignment with the Input Image and the generated Image)

(target_text)
A photo of
two bananas

CLIP Score : 0.4308
SSIM Score : 0.3127

(target_text)
A closed book

CLIP Score : 0.3537
SSIM Score : 0.2621

Discussion and Conclusions

- The major strengths of our approach is that it handles a wide range of edits. It requires only a single image and the text description eliminating the need of masks and multiple views. Further it reserves the key details of the original image, such as textures, background, and object identity.

- There are also few limitations of the model like in some cases, the desired edit is applied very subtly (if at all), therefore not aligning well with the target text. In other cases, the edit is applied well, but it affects extrinsic image details such as zoom or camera angle.

- In totality, by leveraging pre-trained diffusion models, it introduces a simple, unified pipeline capable of complex, photorealistic transformations.

References

- Imagic: Text-Based Real Image Editing with Diffusion Models
  https://arxiv.org/abs/2210.092761
- Image-to-Image Translation with Conditional Adversarial Networks
  [1611.07004] Image-to-Image Translation with Conditional Adversarial Networks
- Unsupervised Image-to-Image Translation Networks
  [1703.00848] Unsupervised Image-to-Image Translation Networks
- Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
  [1703.10593] Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Thank you