Pix2Mask2Pix: CLIP Guided Text-to-Mask-to-Image Editing

Minyang Xie
IIIS
Tsinghua University

Jiayi Hu
IIIS
Tsinghua University

Jiani Wang
IIIS
Tsinghua University

Abstract

Write an abstract, stating that we have achieved a pipeline that can complete the text-to-image editing task by first generating a mask from the text prompt and then inpainting the image from the mask. We also allow for user drawn mask. Compare the pipeline with the existing Pix2Pix model.

1 Introduction

Describe the needs / usages for text guided image editing. Especially the results attained by pix2pix and GPT-4o model series. Then **briefly introduce** the pipeline we have, pointing out the advantages of our pipeline over the existing models (style-keeping, self-drawn mask, etc.). Finally, we shall describe the training work we have done.

2 Related Work

2.1 Pix2Pix

Add related work on Pix2Pix, show pictures, cite the original paper, and describe the model architecture. Also, mention the limitations of Pix2Pix, such as the difficulty in generating complex masks.

2.2 Text guided Image Embedding

Add related work on text guided image editing, such as CLIP, SigCLIP etc. Describe how these models can be used to generate masks from text prompts, and how they can be used to guide the image editing process.

2.3 Text guided Image Inpainting

Cite Stable Diffusion models.

2.4 GPT-4o

Add a photo descibing the GPT-4o image generation tool, and add a few inpainting examples.

3 Pipeline Implementation

3.1 Overview

Plot a flowchart here to show the pipeline we have implemented:

Preprint. Under review.

- User input text prompt and/or mask
- Use CLIP to generate a mask from the text prompt
- Use Qwen to generate a description of the mask
- Use Stable Diffusion to inpaint the image from the mask

3.2 CLIP Mask Generation

Describe how we use CLIP to generate a mask from the text prompt. Show some examples of the masks generated by CLIP.

3.3 Qwen Description

Descibe why we need to use Qwen to generate a description of the mask, and how we use it to guide the image inpainting process. Show some examples of the descriptions generated by Qwen and images generated.

3.4 Stable Diffusion Inpainting

Describe how we use Stable Diffusion to inpaint the image from the mask. Show some examples of the inpainted images.

4 Experiments and Training

We probably need to train the CLIP model on our Magic Brush dataset to improve the mask generation quality. I will upload the training script and how to download the dataset.

We need also think of a metric (maybe miou to verify the mask quality, K-means to verify the mask clustering quality, and FID to verify the image quality) to evaluate the performance of our pipeline.

Also, I have tried to finetune the Pix2Pix model on our Magic Brush dataset, and I will paste a few cherry picked results here.

5 Results

Show a few paired results of our pipeline. You may compare with the results of Pix2Pix and GPT-4o.

6 Ablation Study

6.1 CLIP Mask Generation

Get rid of the CLIP mask generation step, do full-inpainting with Stable Diffusion, and compare the results with the full pipeline.

6.2 Old / New CLIP model

Change the CLIP model to the old one, and compare the results with the full pipeline.

6.3 Qwen Description

Get rid of the Qwen description step, and use the CLIP mask directly to inpaint the image. Compare the results with the full pipeline.

7 Discussion

Some of the limitations of our pipeline, such as the quality of the masks generated by CLIP, the guide strength of the Qwen description, and the quality of the inpainted images. Also, compare with SOTA transformer structured models, note the ability to generate complex masks and drawing text.

8 Conclusion

In this report, we have presented a pipeline that can complete the text-to-image editing task by first generating a mask from the text prompt and then inpainting the image from the mask. We have shown that our pipeline can achieve better results than the existing Pix2Pix model, and we have also allowed for user drawn masks. Also described the training work we have done to improve the mask generation quality.