# COSE474-2024F: Final Project Proposal
# A text-based image editing tool combining CLIP and Stable Diffusion

**2020320041 Seokmin Kim**

## 1. Introduction

Recently, various generative AIs have been performing tasks such as generating or editing images based on user instructions. However, in image editing, generative AI currently provides only limited effects or filters, which often do not reflect the original image's characteristics well. As a result, users still rely on professional software for advanced image editing. Therefore, this project aims to develop an innovative text-based image editing tool by combining CLIP (Contrastive Language-Image Pre-training) and Stable Diffusion. This model will allow users to input editing instructions in natural language, which will be interpreted to modify the original image as desired.

## 2. Problem definition & chanllenges

The challenge is to take the original image provided by the user along with natural language editing instructions and generate a high-quality image that is accurately modified according to those instructions. To achieve this, only specific parts of the image designated by the user must be changed while keeping the rest of the image unchanged. This is implemented by combining the attention mechanism of CLIP with the masking technique of Stable Diffusion. Additionally, the edited parts must harmonize with the overall style of the original image, which is achieved by using the style-transfer function of Stable Diffusion.

## 3. Related Works

CLIP (Radford et al., 2021): Developed by OpenAI, this model learns associations between images and text using a large-scale dataset. In our system, it will be used to understand the user's editing instructions and link them to image features.

Stable Diffusion (Rombach et al., 2022): A latent diffusion model that generates high-quality images based on text descriptions. In our system, it will be adapted to modify specific parts of an existing image.

InstructPix2Pix (Brooks et al., 2022): A model that edits images according to textual instructions, which can serve as an inspiration for the basic architecture of our system.

DiffEdit (Couairon et al., 2022): Proposes a technique for editing specific areas of an image using mask guidance. This approach can be referenced for implementing our precise editing capabilities.

EditGAN (Ling et al., 2021): EditGAN is a generative model designed to perform fine-grained edits on images by manipulating the latent space of a GAN. It offers a high degree of control over image modifications, which can be useful for fine-tuning edits to match user instructions more closely.

DragGAN (Pan et al., 2023): DragGAN allows interactive point-based editing of images, where users can "drag" specific points on an image to a desired position, resulting in highly intuitive and precise modifications. This tool provides an approach for real-time editing that can contribute to the interactive aspects of our system.

## 4. Datasets

For this project, we plan to use the following datasets and, if necessary, construct our own dataset

MS-COCO (Lin et al., 2014): A large-scale image dataset containing various objects and scenes, used to train the model's fundamental image understanding ability.

Conceptual Captions (Sharma et al., 2018): A dataset consisting of image-caption pairs, used to learn the association between text and images.

## 5. State-of-the-art methods and baselines

To evaluate the system's performance, we assess the quality of the generated images using FID (Fréchet Inception Distance) and SSIM (Structural Similarity Index). At this time, we compare the performance with DiffEdit, Imagic, InstructPix2Pix, EditGAN, and DragGAN.

## 6. Schedule & Roles

Weeks 1: Data collection and preprocessing, basic model implementation

Weeks 2: Integration and training of CLIP and Stable Diffusion

Weeks 3: Improvement of editing accuracy

Weeks 4: Performance evaluation, result analysis, and report writing

## References

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23556–23565, 2022.

Couairon, G., Grechka, A., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9213–9223, 2022.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, Cham, 2014.

Ling, H., Zou, K., Zhang, Z., Sun, C., and Jia, J. Editgan: High-precision editing with gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1211–1220, 2021.

Pan, P., Ma, L., Zhang, Y., Li, Z., Fan, C., Li, Y., and Smith, A. Draggan: Extending interactive point-based editing for gan images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5402–5411, 2023.

Radford, A., Kim, J. H., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.