Text-Driven Image-to-Image Generation

Amir Moris
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
amir.moris35@gmail.com

Verina Gad
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
verinagad258@gmail.com

Maria Tawfek
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
mariatawfek388@gmail.com

Mina Girgis
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
minagirgis2001@gmail.com

Carolina George
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
carolinageorge510@gmail.com

Yomna A. Kawashti
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
yomna.ahmed@cis.asu.edu.eg

Ghada Elnahas
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
ghadaelnahas@cis.asu.edu.eg

Hanan Hindy
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
hanan.hindy@cis.asu.edu.eg

Abdel-Badeeh Salem
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
absalem@cis.asu.edu.eg

Figure 1: Samples of InstructPix2Pix model output that takes an input image and edit prompt, to generate output image with the edit instruction while preserving the object characteristics.

***Abstract*—** This work aims to apply a deep learning model to update images based on text input, essential for social media, advertising, and content creation. Users can upload an image and provide text instructions for modifications. The system interprets these instructions to update the image while maintaining high visual quality.

Key features include an intuitive user interface, advanced text interpretation algorithms, and the ability to change visual elements as needed. Experiments revealed high memory and GPU demands. Despite this, two models emerged as both efficient and lightweight. These two models were SINE and InstructPix2Pix. Initial experiments involved the SINE model, which comprises two main parts: training the text-to-image diffusion model on a dataset and fine-tuning it on the input image. Although more lightweight, the model still demanded extensive resources. Recognizing the need for a more resource-efficient solution, the InstructPix2Pix model was ultimately utilized, demonstrating efficiency in both performance and memory usage.

A challenge was the lack of domain-specific datasets. A domain-specific dataset focusing on "products" was created using a cost-

effective, hardcoded text synthesizer instead of large language models (LLMs) such as GPT-3. This provided high-quality text prompts for the dataset.

In conclusion, this project offers a simple and effective way to update images using text instructions and introduces a specialized products dataset, which is uncommon in the field. It benefits content creators, educators, marketers, and social media users by helping them create customized and engaging visual content.[1]

I. INTRODUCTION

In recent years, the demand for high-quality, domain-specific image generation has become increasingly critical. Existing image generation models often lack the precision required for specific industries, particularly in e-commerce and marketing, where the accuracy and quality of product images are essential. This project addresses these challenges by focusing on the advancement of text-driven image-to-image generation technology tailored for the product domain.

The theoretical foundation builds on recent advancements in image generation models, which can be broadly categorized into four types based on input, output, method, or architecture: segmentation mask image-to-image generation, text-to-image generation, image-to-image generation, and text-driven image-to-image generation.

Segmentation mask image-to-image generation involves creating images from segmentation masks while incorporating textual descriptions. This approach combines text and segmentation masks as inputs to produce realistic images that align with the provided descriptions. **Text-to-image generation models**, on the other hand, generate high-quality images solely from user-provided text prompts. Recent diffusion-based models have excelled in this area by conducting the diffusion process in latent space, which reduces sampling steps without compromising image quality.

Image-to-image generation enables the manipulation and transformation of existing images using deep learning architectures. These models map an input image from a source domain to a target domain, resulting in a semantically equivalent image with the desired alterations. Lastly**, text-driven image-to-image generatio**n leverages large text-to-image models to synthesize images based on a text prompt. Some models use multiple input images and an edit instruction, while others use a single input image and a text prompt to generate the output image.

Despite advancements in image generation models, there remains a significant gap in the availability of domain-specific datasets, particularly for product images. Existing models often lack the domain-specific focus necessary to achieve high accuracies and quality in the generated images. This is a critical issue for applications, where precise and high-quality product images are essential for user engagement and sales.

Our research aims to enhance the generation of new images based on a guidance image and a text prompt, specifically targeting product images. Unlike many existing solutions, our approach concentrated on generating a custom dataset that meets the specific needs of the product domain. This specialization is crucial for applications in e-commerce, marketing, and inventory management, where the fidelity and contextual relevance of images directly impact user engagement and sales.

To address the dataset gap, we implemented a comprehensive process for generating a robust and domain-specific dataset tailored to product images. This process involved two main components: textual data generation and image data generation. For Textual Data Generation, we implemented a hardcoded text synthesizer to produce high-quality prompts. This approach, while simpler than utilizing large language models such as GPT-3, proved cost-effective and efficient. The synthesizer generated diverse and contextually relevant prompts that precisely described modifications needed for product images, ensuring the textual data's precision and relevance to the product domain. Subsequently, using these generated text prompts, we created corresponding images by selecting or creating base product images and modifying them according to the textual descriptions. These modifications encompassed adjustments in color, texture, perspective, and other specified attributes, forming a crucial step in constructing a dataset that faithfully represents the required variations and specifics in product imagery.

We utilized cutting-edge deep learning models and sophisticated algorithms, with a particular emphasis on diffusion models and stable diffusion techniques. These models leverage iterative noise manipulation in latent space to produce high-quality, realistic outputs. Additionally, Conditional Generative Adversarial Networks (GANs) play a pivotal role in achieving accurate transformations based on textual cues, ensuring the generated images align closely with the provided descriptions.

In summary, our research provides a robust framework for automating the generation and editing of product images through advanced text-driven models. By creating a specialized dataset tailored to the product domain, we ensure that the generated images are of high quality and contextually accurate, thereby enhancing their utility in various industry applications.

II. RELATED WORK

Text-Driven Image-to-Image Generation is a computer vision task that involves transforming an existing image based on textual descriptions to modify or enhance its content.

Text-driven image-to-image generation recently developed large text-to-image models have shown unprecedented capabilities, by enabling high-quality and diverse synthesis of images based on a text prompt written in natural language. Some models [5] enable such tasks by taking more than one input image and an edit instruction to generate an output image for different tasks like art rendition, view synthesis, and property modification. Other models [1, 4, 7] enable generating output image by using only one input image provided by the user in addition to the edit instruction, Also there are subject-driven models [5, 8, 9] that take one input image and subject text to generate the output image.

Significant papers that are related to our field of study will be reviewed.

Ruiz et al. [5] proposed "DreamBooth". They addressed in their paper a key limitation in text-to-image generation: Traditional models often struggle to accurately render specific subjects or objects that are not well-represented in their training data. Alternatively, DreamBooth provides the ability to customize and fine-tune models to generate images of particular subjects based on minimal data through a fine-tuning technique for text-to-image diffusion models that allows for precise and personalized generation of images.

Despite its advancements, DreamBooth has certain limitations. One primary limitation is the dependency on a small set of images for fine-tuning. While this reduces the need for large datasets, it may still pose challenges when the available images are not representative enough of the subject's variability. Additionally, the fine-tuning process can be computationally expensive and time-consuming, potentially limiting its applicability in real-time or resource-constrained environments.

Zhang et al. [4] proposed "SINE" which addresses different challenge of editing specific attributes of a single image using textual descriptions, leveraging the capabilities of diffusion models to achieve high-fidelity edits. This approach is particularly useful for applications where only a single image of the subject is available, and exact text-guided modifications are required. Building upon the foundational capabilities of diffusion models, SINE employs

[1] You can check our project on hugging face and our generated dataset

techniques such as Classifier-Free Guidance (CFG) and latent code optimization to iteratively refine the image based on the given textual instructions.

The key innovation in SINE lies in its ability to focus edits on specific regions of the image while preserving the overall structure and context. This is facilitated through diffusion model methodologies that utilize attention mechanisms to focus on specific parts of an image during the editing process, and region-specific loss functions to ensure that changes made to targeted areas align with the intended modifications described in textual input.

Despite its strengths, SINE faces limitations such as the need for fine-tuning, which demands substantial computational resources and time. The method's effectiveness is sensitive to the choice of guidance steps and weights and the quality of the edited images depends significantly on the initial input image and fine-tuning data.

Hertz et al [6] proposed "Prompt-to-Prompt (P2P)" which uses a different approach which is word-swap. The idea is to leverage the "cross-attention" mechanism in text-to-image models, which connects the textual prompt to different parts of the image during generation. As it takes as input the input image, input text that describes the input image, and output text that describes the desired output image. Using the word-swap technique, it learns which attention map in the input image to change or manipulate.

Brooks et al [1]. proposed "InstructPix2Pix" The distinctiveness of the Pix2Pix approach lies in its conditioning mechanism, where both the generator and discriminator are conditioned on the input image, ensuring that the generated output is not only realistic but also contextually aligned with the input. This differentiates Pix2Pix from traditional (GANs), which typically generate images from random noise. The model is trained on an AI-generated paired dataset, comprising input images and their corresponding target images, allowing it to learn a direct mapping from input to output. Such an approach is particularly effective in tasks such as transforming sketches into photographs and colorizing grayscale images, where the structured relationship between the input and output pairs is leveraged to produce high-fidelity translations.

Pix2Pix introduces an innovative approach to dataset creation by using GPT-3 to generate textual datasets. This includes input image descriptions, edit instructions, and output image descriptions. A stable diffusion model is then used to generate the input image based on these descriptions.

Using the input image and its description, Pix2Pix employs P2P model to generate the output image, effectively translating textual descriptions into visual modifications.

InstructPix2Pix, despite its strengths, exhibits limitations. AI-generated training data can introduce biases and restrict the model's ability to handle unseen concepts. Natural language ambiguity may lead to misinterpretations of editing instructions. The model may struggle with highly specific edits or complex scene manipulations due to semantic understanding limitations.

Table 1 shows a comparison between the discussed papers and provides a summary of the techniques and datasets used.

Table 1: Comparison between papers

Paper	**Technique**	**Dataset**
DreamBooth [5]	Generative Diffusion models (Imagen - Stable diffusion model) 1. Fine-tuning 2. class-specific prior-preservation loss	MS COCO LAION-Aesthetics V2 6.5+
SINE [4]	The linear combination function operates between a text-to-image diffusion model trained on a dataset and a diffusion model trained on user-entered inputs.	LAION dataset flickr.com unsplash.com
Prompt-to-Prompt [6]	Image editing with cross-attention control and word swap technique.	COCO Dataset PIE-Bench
InstructPix2Pix [1]	GANs & fine-tuning a text-to-image diffusion model.	Ai-Generated

III. METHOD

Following InstructPix2Pix [1]. First, we generate our domain-specified dataset. Second, we use a pre-trained text-to-image diffusion model on a large-scale dataset that varies in classes. Despite being trained with Ai-generated images, the pre-trained model is able to generalize to editing real images.

The pre-trained model was trained on a an AI-generated dataset which is used to train a conditional diffusion model that edits images from written instructions. The main model is based on Stable Diffusion, a large-scale text-to-image latent diffusion model [3]. Latent diffusion improves the efficiency and quality of diffusion models by operating in the latent space of a pre-trained variational autoencoder with encoder and decoder.

The text-to-image model utilizes the stable diffusion model shown in Figure 2, which is trained using the data. Once trained, the model can take the input image and edit instruction and generate a new image with the desired modifications. This part is crucial for interpreting user instructions and transforming them into visual content.

The diffusion model in latent space with conditioning involves encoding the input image and text instructions, introducing noise through a diffusion process, and then iteratively denoising the latent representation with guidance from the text instructions. The text encoder processes the edit instructions and provides the necessary conditioning information. The denoising process is carried out by the U-Net, which efficiently removes noise and refines the latent representation with a conditional input for the input image and the edit instruction that is passed to the Architecture of the U-Net as an encoded text embedding, a timestamp which is also encoded. The original U-Net architecture is to generate the output image from a random noise but after adding two more conditional inputs which are the input image and the edit instruction, the final clean latent representation is decoded back into an image that reflects the user-specified edits. The cross-attention mechanism plays a crucial role in integrating textual information during the denoising process, ensuring the edits are applied effectively and harmoniously.

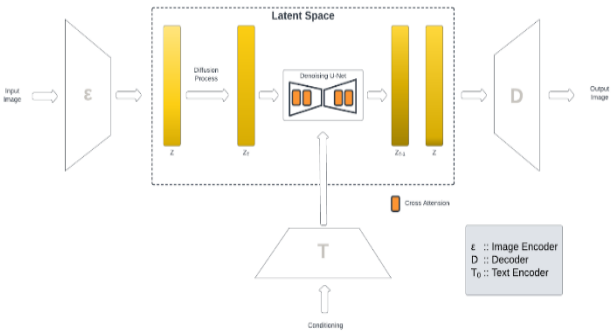

Figure 2: Diffusion Model

Lastly to maintain the impact of the edit instruction and the input image on the output image, we use Classifier-Free Guidance [2], which is a method that balances the influence of the edit text instruction and input image feature. This mechanism ensures that the generated image accurately reflects the user's specifications, maintaining a harmonious integration of the text and image effects. This balance is essential for producing coherent and visually appealing results, Figure 3 shows an illustration of the effect of classifier free guidance for using different values for edit instruction and input image on the output image.

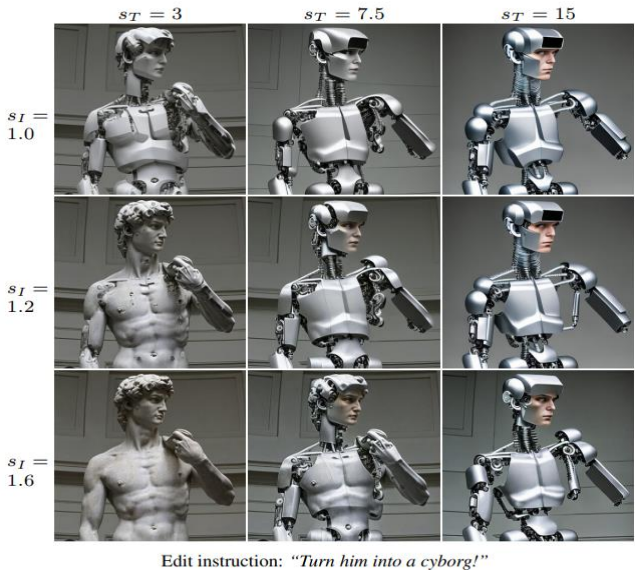

Figure 3: Classifier-free guidance weights over two conditional inputs. S_I controls similarity with the input image, while S_T controls consistency with the edit instruction [1].

The training process for this architecture in Figure 2 requires significant resources to train the model, so we couldn`t train the model due to the lack of resources.

IV. DATASETS

This paper focuses on the generation of a domain-specified dataset, as we noticed a lack of availability for domain-specific datasets in the Text-driven Image-To-Image field, and we chose "products" to be our domain of interest. The dataset generation process contains two sub-processes:

A- Textual Data Generation:

Initially, the textual dataset generation was attempted using a large language model (LLM) such as GPT-3. However, due to the high associated costs, we adopted an alternative approach: a hardcoded text synthesizer.

We utilize a hardcoded text synthesizer to scrape Unsplash[2] for image descriptions of various products, categorized into fashion, electronics, furniture, etc. We apply different editing instructions to these descriptions, enabling approximately 20 editing effects on the input image descriptions including color, weather, material, style, seasons and other aspects.

	Manual Dataset	Automated Dataset
Number of categories	50	10
Number of different edits	50	25
Total number of rows	50	3000
Number of images for each row	6	6-10
Final size	~300	~12 000

Table 2: Dataset Summarization

This synthesizer accepts a JSON file containing randomly generated input texts and maps these texts to various edit instructions. The process outputs a JSON file that includes the input text, the corresponding edit instruction, and the resultant output text. Despite being considered a naive approach for textual data generation, the hardcoded text synthesizer produced high-quality text prompts, proving successful in creating the required dataset, As shown in Table 2, a statistical summary of our generated dataset.

B- Image Data Generation:

We use the input image description to generate the input image itself by utilizing a text-to-image diffusion model. Then we use Prompt-to-Prompt [6] to generate the output image, so for each row in the textual Data Generation process, we can obtain multiple rows in the Image Data generation as the diffusion model and Prompt-to-Prompt [6] enables us to generate different pairs of input image and output image.

Figure 4 shows the complete dataset generation pipeline.

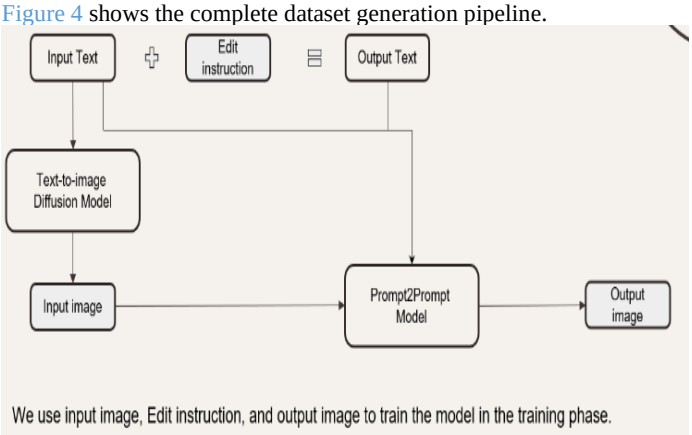

Figure 4: Dataset Generation

[2] Unsplash website

By combining process A with process B, we end up obtaining 5 columns which are:

- Input image description: used to generate input image.
- Edit instruction: used to generate output image description and used in the training process.
- Output image description: used to generate the output image.
- Input image: used to generate output image and used in the training process.
- Output image: used in the training process as the ground truth of the Image-To-Image model.

However, the model only needs input image, edit instruction, and output image in the process of fine-tuning.

Figure 5 illustrates various input images along with their text descriptions, the applied edit instructions, and the resulting output images.

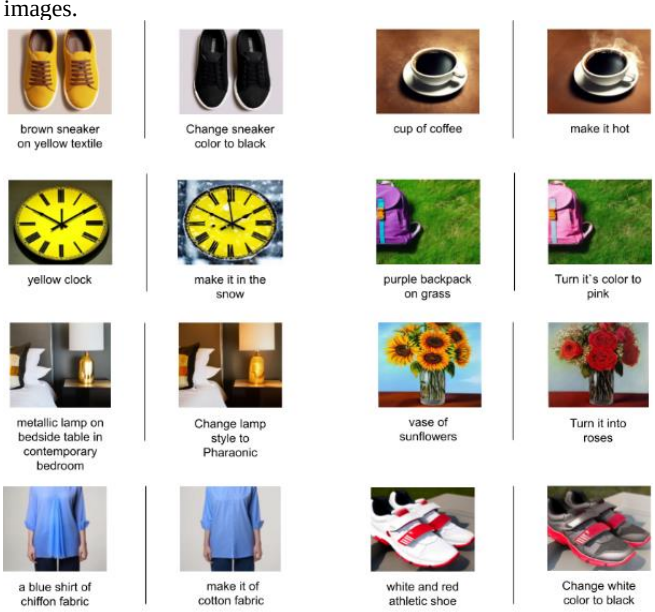

Figure 5: Generated Dataset Samples

V. RESULTS

In this section, the various experiments and results applied throughout the work on our project will be presented.

We initially worked on the Dreambooth model [5], but it uses inputs different from our architecture. Specifically, Dreambooth requires 3-5 input images, while we aimed to use only one input image. Then we tried using the Sine model [4], the sine model consists of two main parts in its architecture. The first part is training the text-to-image diffusion model on the dataset, and the second part is fine-tuning the text-to-image diffusion model on the input image. The output image is generated by using a linear combination function for both parts. The architecture of the SINE model involves training a text-to-image diffusion model on a dataset, followed by fine-tuning the model on the specific input image. The final output is generated by combining the results from both the fine-tuned diffusion model and the original diffusion model using a linear combination function. This approach helps produce images that align with the text description while preserving the input image's characteristics. However, the fine-tuned text-to-image diffusion model requires substantial resources, making it challenging to fully utilize the architecture. Despite this, we were able to generate output images by omitting the fine-tuning part, though the results were less accurate and might differ from the input image. Due to resource limitations that prevented the SINE model architecture from preserving object features, we transitioned to the InstructPix2Pix model [1]. This model was chosen for its efficiency in performance and memory usage. However, it required the creation of a product-based dataset specifically designed to fine-tune the InstructPix2Pix model. Initially, the textual dataset generation was attempted using a large language model (LLM) such as GPT-3. However, due to the high associated costs, we adopted an alternative approach: a hardcoded text synthesizer. This synthesizer accepts a JSON file containing randomly generated input texts and maps these texts to various edit instructions related to attributes such as color, weather, material, style, and seasons. The process outputs a JSON file that includes the input text, the corresponding edit instruction, and the resultant output text. Despite being considered a naive approach for textual data generation, the hardcoded text synthesizer produced high-quality text prompts, proving successful in creating the required dataset. The training process for InstructPix2Pix requires substantial memory and GPU resources. the minimum resources needed include 64 GB of VRAM. To address this, we attempted to minimize the parameters and partition the dataset during the training process to enable fine-tuning. However, the resources provided by Kaggle were still insufficient. This trial involved testing various platforms offering free resources, including Google Cloud and Microsoft Azure, which supported only CPUs, all of which proved unsuitable except Kaggle. Consequently, Kaggle was utilized for both dataset generation and model deployment. Figure 6 illustrates a representative instance of the model being employed on Kaggle.

Figure 6: Example

The model itself is deployed on Kaggle, but the user interface is on Hugging Face. The process of generating the output image follows the steps shown in *Figure 7*. These steps can be summarized as follows:

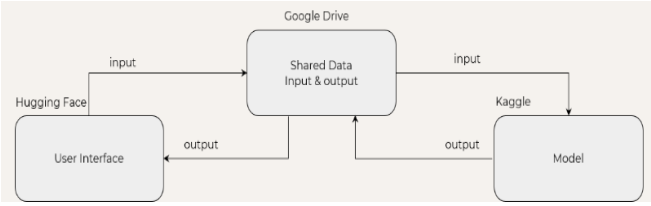

Figure 7: Deployment

a. The user enters the input on Hugging Face, then the input is uploaded to Google Drive using the Google Drive API.
b. A trigger is then called using Kaggle-API to inform the notebook on Kaggle to start running. Once the notebook starts running, it first installs the essential dependencies, which takes approximately 20 minutes. It then retrieves the input from Google Drive, runs the model to generate the output image, and uploads the image to Google Drive.
c. On Hugging Face, we continuously check if the notebook on Kaggle has been completed and run successfully. If so, we retrieve the output image from Google Drive and display it to the user.

VI. CONCLUSION

In summary, this project aimed to develop a system for text-driven image-to-image generation, leveraging advanced deep learning models and algorithms to automate the process of updating images based on text inputs. The project involved several key components:

(1) System Development: We created a user-friendly interface that allows users to upload an image, provide text instructions for modifications, and generate updated images. The system utilizes the pretrained text-to-image diffusion model. The process is designed to maintain high visual quality while accurately applying the textual modifications. The user interface was hosted on Hugging Face.

(2) Dataset Generation: A significant part of the project was generating a suitable dataset. Initially, we explored using large language models like GPT-3 for text prompt generation but opted for a more resource-efficient approach using a hardcoded text synthesizer. The generated text prompts were high quality and suitable for creating the necessary dataset. The dataset generation and model deployment were facilitated using Kaggle. Figure 5 illustrates various input images along with their text descriptions, the applied edit instructions, and the resulting output images.

REFERENCES

[1] T. Brooks, A. Holynski, and A. A. Efros "InstructPix2Pix: Learning to Follow Image Editing Instructions" in University of California, Berkeley, 2022.

[2] J. Ho and T. Salimans "CLASSIFIER-FREE DIFFUSION GUIDANCE " in Google Research, Brain team, 2022.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer "High-Resolution Image Synthesis with Latent Diffusion Models" in Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany, 2021

[4] Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren "SINE: SINgle Image Editing with Text-to-Image Diffusion Models" in Rutgers University, Snap Inc, 2022.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”, in Boston University, 2023.

[6] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman1, Y. Pritch, and D. Cohen, "Prompt-to-Prompt Image Editing with Cross Attention Control", Google Research, The Blavatnik School of Computer Science, Tel Aviv University, 2022

[7] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel "Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation" in Weizmann Institute of Science, 2022.

[8] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, and W. Zhu "DisenBooth: Identity-Preserving Disentangled Tuning for Subject-Driven Text-to-Image Generation" in Department of Computer Science and Technology, Tsinghua University, Beijing National Resear

[9] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani "Imagic: Text-Based Real Image Editing with Diffusion Models" in Weizmann Institute of Science, 2022.