Xi'an Jiaotong-Liverpool University

西交利物浦大学

DTS311TC Final Year Project

Text-driven Image Editing in Diffusion Model with Attention Control Methods

In Partial Fulfillment of
the Requirements for the Degree of
Bachelor of Engineering

By
Haosen Zhong
ID 2035643

Supervisor Name
Dr. Xianxu Hou

School of AI and Advanced Computing
XI'AN JIAOTONG-LIVERPOOL UNIVERSITY

April 2024

Abstract

In recent years, significant attention has been directed towards large-scale text-driven diffusion models, recognized for their capacity to generate diverse and high-fidelity images guided by text prompts. Despite their remarkable image generation capabilities, they are not suitable for image editing because any modification of the text prompts result in entirely distinct output images when using diffusion models. While recent researches have explored approaches involving masks to localize edited image regions, they overlook the fundamental issue of structural incoherence between the mask region and the image content. To address the existing problems in image editing, the primary aim of this research is to achieve text-driven image editing method which can preserve the original image's structure. This will be achieved through precise control over the cross-attention map and its integration into the diffusion process.

This research project undertook a comprehensive analysis of the determinants influencing spatial features during the diffusion process, that is, the cross-attention map and the text token vector values jointly. This study tried to control the spatial features of the edited image by controlling the cross-attention map corresponding to the edited text in the forward process of the diffusion model. Employing a stable diffusion model, this study conducted experiments to refine the control method iteratively. Ultimately, three distinct editing effects were achieved. First, users could replace words in the text to effect changes in elements within the image. Second, users could augment the original text with additional words, enriching the image with diverse elements and styles. Third, users could re-weight the semantic impact of words and phrases within the text, influencing the emotional response portrayed in the edited image. This approach only requires control of the cross-attention map in the forward process, obviating the need for model retraining, which significantly reduces the time required for image editing while concurrently ensuring a certain level of editing precision and fidelity to the original text.

Key Words: text-to-image generation, image editing, diffusion model, stable diffusion, cross-attention maps, spatial feature

Acknowledgements

First of all, for this final year project about image editing based on diffusion model in the field of text-generated images, I really want to express my special thanks to Professor Xianxu Hou for his careful guidance and help. In the early stage of my senior year, Professor Hou taught me various knowledge about diffusion model, which helped me to have a better understanding and learning in this field. In the middle and late stages, I am very grateful to Professor Hou for answering my questions in detail. It is very lucky for me to meet Mr.Hou.

Secondly, I would like to thank my teammates and good friends during the academic period, the mutual help among friends makes me enthusiastic about the future and positively solves the difficulties that may arise. I am very grateful for the company of my good friends during the four years of university.

Finally, I would like to thank Xi'an Jiaotong-Liverpool University for offering us this subject at the end of the four-year undergraduate period, which is undoubtedly the best test of our four-year study. Xi'an Jiaotong-Liverpool University has provided us with a free learning environment, which has fostered my ability to study independently. I am constantly able to see the results of my learning and motivated to improve myself.

Contents

List of Figures

Chapter 1

Introduction

In recent years, text-to-image generation techniques have made significant progress in the field of artificial intelligence. In particular, large-scale language-image (LLI) models based on diffusion model and auto-regression, including Stable Diffusion (SD) [1], RAPHAEL [2], GLIDE [3], OpenAI's DALL-E [4], Google's Imagen [5] and Parti [6], have been widely used for generating high-quality images by providing text prompts. These models can incrementally add noise to the image and learn the inverse process based on complex text prompts, thus generating various types of realistic images. However, although these models perform very well in generating images, they have different shortcomings in terms of accurate editing of images. The current text-to-image models derived from diffusion-based model do not focus on image editing and lack control over specific regions of the image. This results in their sensitivity to adjustments in text prompts, and even very small changes in may cause the model to produce an image that is completely different from the original schematic. Users cannot directly use diffusion model to generate images that have been locally edited in response to changes in text prompts. This can be understood easily by the image explanation in Fig.1.1. These limitations motivate the need for accurate editing of image content driven by only changing the text prompt.

1.1 Motivation, aims and objective

To attain precise image editing, LLI-based methods [3, 7] have introduced editing techniques that exclusively manipulate the image area designated by the user's masks. While this strategy effectively preserves the content of the unaltered portion of the image, it inadvertently sacrifices the original structural information within the masked region. Consequently, this loss of detailed information often leads to image distortion and falls short

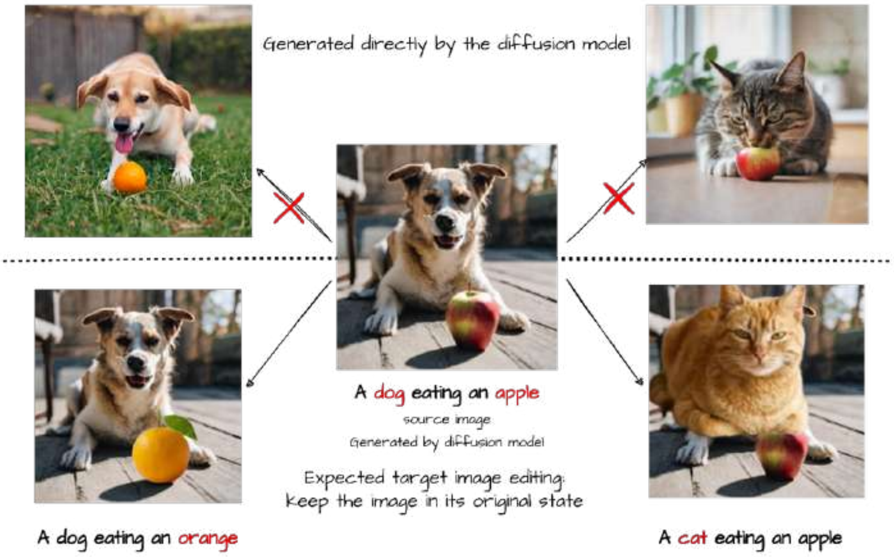

Generated directly by the diffusion model
A dog eating an apple
source image
Generated by diffusion model
Expected target image editing:
keep the image in its original state
A dog eating an orange
A cat eating an apple

Fig. 1.1: Expected results. Center: generate the image using the text prompt [A dog eating an apple]. Top: change the word from dog and apple to cat and orange, then generate the image directly through stable diffusion model. Bottom: use text-driven image editing framework to get the expected result which is only the shape of dog and apple will change to cat and orange and other objects' structure and shape will be preserved.

of achieving the desired level of editing precision. Therefore, while existing LLI models have made considerable strides in image generation, their applicability to precise image editing remains constrained. Presently, numerous models are under investigation for image editing purposes, yet they exhibit varying degrees of limitations concerning accuracy, structural adaptability, and computational efficiency.

In this paper, to achieve precise editing of text-guided images, this study delves into controlling the image generation process of the diffusion model post text editing. In order to maintain content and structural stability of the image before and after editing, this study extensively analyzes the determinants of spatial features within the diffusion model. To this end, I conduct an in-depth examination of the relationship between the attention map and the text token, and their influence on the spatial features of the image within the cross-attention layer. Specifically, this study explores how the cross-attention map and the text token collaborate to determine the spatial features of the generated image.

Since the cross-attention map is a key element in controlling the spatial layout and geometry of an image, this study can consider controlling the cross-attention maps in the diffusion process after text editing. The cross-attention map acts as a high level tensor that can combine pixels and tokens. Our key idea is to control the spatial features of the image by injecting the cross-attention maps of the source image during the diffusion process after the text has been edited, and to control the time step and the degree of constraint of the attention map injection according to different editing requirements. Based on this idea, I wrote an editing framework for controlling the attention maps based on text editing, and implemented the image editing task for three text editing modalities. In Fig.1.2, I show the images before and after text editing corresponding to these three ways. The first is when the user wants to edit a token in the text (e.g., dog to cat), the method will fixing the attention map of the dog and injecting it into the image generation of the cat to preserve the scene structure composition of the image. The second method is when the user adds a new phrase to the source text, the method will freeze the attention map of the tokens common to both texts and inject them into the edited diffusion process, allowing the new attention to flow to the new token to diversify the editing. The third method is when the user wants to re-weight (strengthen or weaken) the semantic effect of a word in the image, the method will set parameters to scale the size of the attention map.

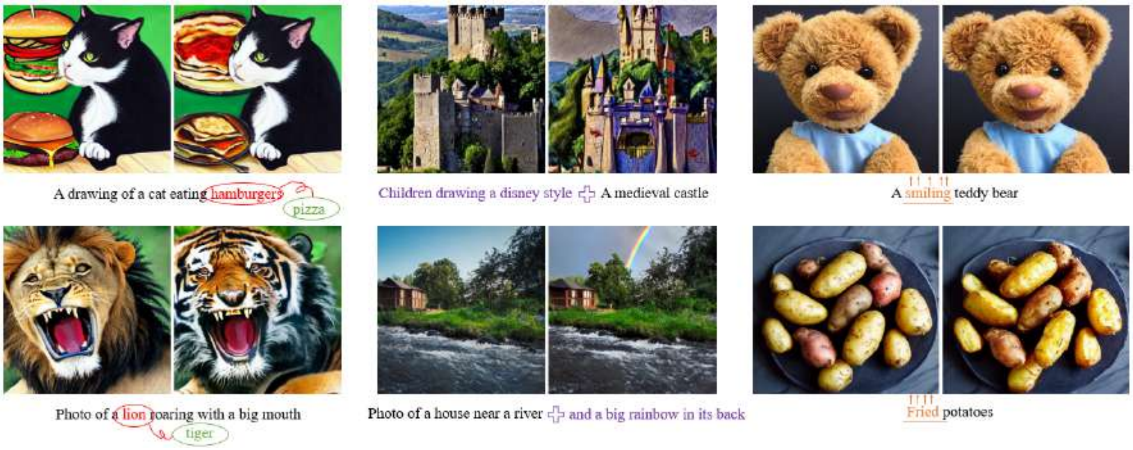

A drawing of a cat eating hamburgers
pizza
Children drawing a disney style ✚ A medieval castle
A smiling teddy bear
Photo of a lion roaring with a big mouth
tiger
Photo of a house near a river ✚ and a big rainbow in its back
Fried potatoes

Fig. 1.2: A preview of editing results comparison. In each group of images, left shows the original image generated by the stable diffusion model and right shows the edited image using the different methods. *left 2 figures*: Text replacement; *middle 2 figures*: Text refinement; *right 2 figures*: Text semantic re-weighting.

In short, the algorithms and functions within the editing framework achieve three different effects of image editing and it can be called more professional as text-driven image editing. This method does not require image feature extraction and model training during the editing task like AI face-swapping, so users can get results more quickly when using this method for image editing. Throughout the research process, I also recognised the relationship between the fidelity of the edited text cues and the source image and tried to reduce the distortion of the edited image. Through constant adaptation, text-driven-editing can perform image editing to a greater extent, and in this paper I demonstrate the feasibility of the method with a number of examples.

To sum up, my main contributions in this work about the image editing are as follows:

- This study analyses the relationship between cross-attention and the spatial features of the generated image in depth for the diffusion model, where the cross-attention map and the text token vector values jointly determine the spatial features of the image, and this study considers the control of the cross-attention map to achieve the project objectives

- This study considers controlling the cross-attention maps of the edited image through three methods to achieve the project objectives, namely text replacement, text refinement and text semantic re-weighting.

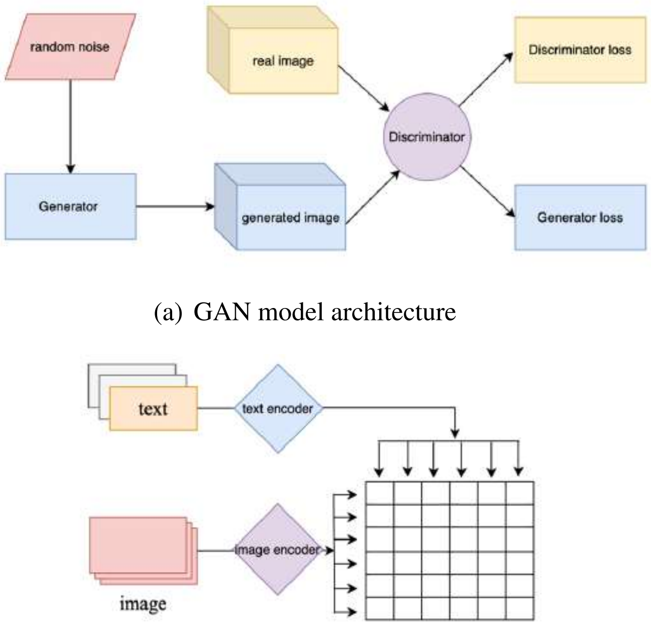
- This study uses the stable diffusion model to conduct experiments, through a large number of experiments to observe the generation and editing of images. The experimental results show that text replacement and text refinement have better results, and most of the time they are able to generate satisfactory edited images.

1.2 Literature review

1.2.1 Text-to-image generation with generative adversarial network

Generative adversarial network. With the rise of deep learning and artificial intelligence, image editing has evolved from the point where users can manually edit images on digital and software-based editing software (e.g., Adobe Photoshop [8], GIMP [9]) to the point where advanced editing effects can be achieved by non-professional users directly through AI. The fields of image recognition and image synthesis have continued to move into the realm of AI research, starting with AlexNet [10] demonstration of the performance benefits of deep convolutional neural networks on large-scale image recognition tasks. Early Generative adversarial network (GAN) [11] introduced game theory-based learning methods and trained on datasets to generate face samples, from 2015 to 2020 GANs was applied to various image processing tasks such as image restoration, super resolution and denoising tasks. Subsequently NVIDIA completely improved the GANs architecture and used StyleGAN2-ada [12] to generate images that are indistinguishable from natural images. The development of GAN has provided a powerful energy for high quality image generation. This can be shown in Fig.1.3.

The emergence of the Transformer architecture between 2020 and 2022 has led to even more groundbreaking developments in image synthesis, and the emergence of CLIP [13] has enabled the combination of CLIP and GAN and helped users to achieve complex image editing through textual descriptions alone. CLIP, as a multimodal neural network, can process both image and text data, and the combination of the two improves the model's understanding and generation of both images and text. However, due to the nature of



(a) GAN model architecture

(b) CLIP model architecture

Fig. 1.3: Model architecture comparison between GAN and CLIP

GANs [14], the combination of the two is not able to achieve better text-driven image editing on diverse datasets, even though it is very good at face processing. In order to achieve more accurate and expressive editing of high-resolution images, Crowson et al. [15] experimented on different datasets using VQ-GAN [16] and various diffusion models, which often outperform GANs that focuses on face generation, using user-supplied masks as a guide to apply and modify editing in the corresponding region of the mask, but these operations need to be performed before each training session. However, these operations need to be fed into the training network before each training session, which is not only cumbersome but also may ignore the original structural information of the image. Accurate text-driven image editing needs to ensure that when the user modifies the target region, the rest of the image is preserved to the greatest extent possible and matches the original image structure. The mask-based modifications described above do not achieve this.

1.2.2 Text-to-image generation with diffusion models

Diffusion models. In 2021 diffusion models [7, 17, 18] introduced a different approach to image synthesis from GANs, where researchers perform machine learning by reconstructing images from artificially added noise [19], this some series of image generation conditioned on plain text is collectively referred to as text image synthesis. Currently large-scale language-image (LLI) models based on diffusion model and auto-regression, including Stable Diffusion (SD) [1], RAPHAEL [2], GLIDE [3], OpenAI's DALL-E [4], Google's Imagen [5] and Parti [6] have been widely used to generate high-quality images. These models can generate a wide variety of high-quality images based on user-supplied text. The researchers wanted to test whether they could be applied to image editing, but found that the models are very sensitive to adjustments to textual prompts, and that even very small changes in the textual prompts can cause the models to produce images that are completely different from the original schematic [20]. These models benefit from large training datasets owned by high-tech companies, and are generated by randomly filtering images from a database and generating images that meet textual constraints, so users cannot accurately edit an image by just changing the text. These limitations have fueled the need for precise editing of image content.

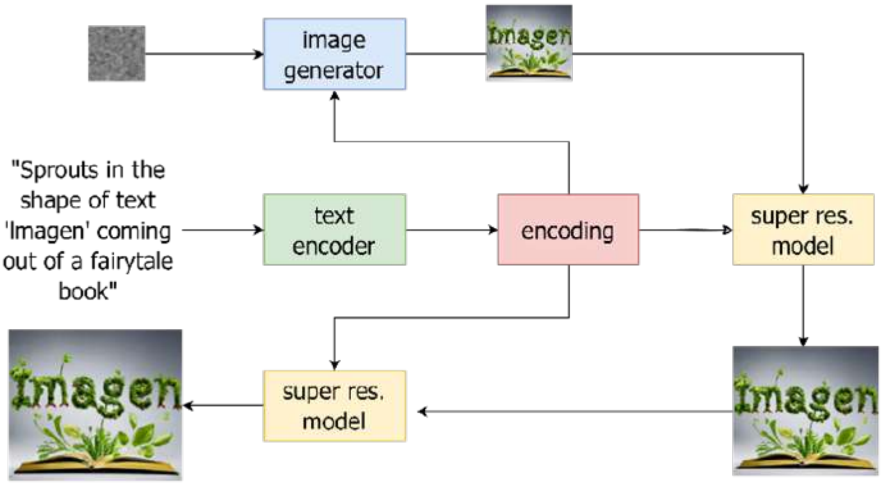


Fig. 1.4: An usual diffusion model architecture

Stable diffusion. Among the myriad diffusion models, the stable diffusion model has garnered significant attention for its ability to progressively unveil details and produce high-quality images. This model leverages the Latent Diffusion Model (LDM) [1], wherein the training objective involves the continuous application of Gaussian noise to training images, akin to a series of denoising autoencoders.

The stable diffusion model comprises three key components: the Variational Autoencoder (VAE) [21], U-Net [22], and an optional text encoder. The VAE encoder compresses the image from pixel space to a lower-dimensional latent space, capturing the underlying semantic meaning of the image. Subsequently, during the forward diffusion process, Gaussian noise is iteratively applied to the compressed latent representation.

The U-Net block, which incorporates a ResNet [23] backbone, then denoises the output of the forward diffusion backward to obtain a latent representation. Notably, this denoising step can be conditioned on various factors such as text strings, images, or other forms. Through a cross-attention mechanism, the encoded conditioning data interacts with the denoising U-Nets. In the case of text conditioning, a fixed, pretrained CLIP ViT-L/14 text encoder is employed to convert text prompts into an embedding space.

Finally, the VAE decoder reconstructs the final image by converting the latent representation back into pixel space. The overall architecture of the stable diffusion model can be viewed as Fig.1.5.

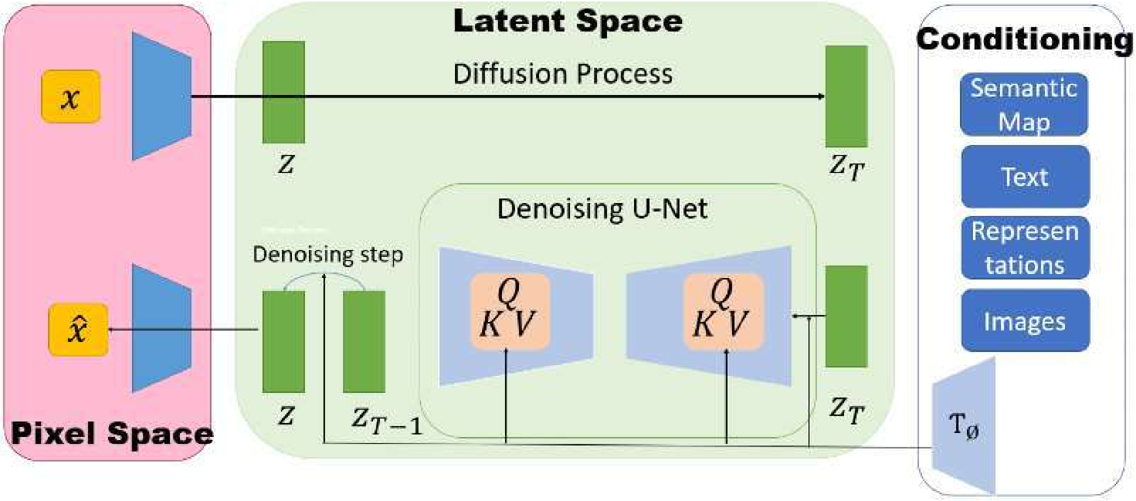


Fig. 1.5: Stable diffusion (SD) model architecture

1.2.3 Cross-attention layer in diffusion models

In the realm of Artificial Intelligence, the introduction of the cross-attention mechanism in Transformer architecture has marked a significant leap forward. This innovative structure revolutionizes the conventional architecture of recurrent neural networks, substantially enhancing the efficacy of natural language processing and various other machine learning tasks. The cross-attention mechanism, a novel approach, involves weighting different positions within the input sequence to extract pivotal information via a self-attention mechanism. Unlike recurrent neural networks, Transformer [24] diverges from sequential processing, instead focusing on all positions simultaneously through self-attention, thereby drastically reducing the time complexity of handling lengthy sequences. Furthermore, Transformer harnesses a multi-head self-attention mechanism to capture intricate semantic relationships across input and output sequences.

In the context of diffusion models, the cross-attention layer plays a vital role by evaluating the similarity between the query sequence and the key-value pair sequence. It utilizes these similarities to weigh the aggregated value sequences, facilitating efficient interaction of correlated information between disparate sequences. The incorporation of cross-attention facilitates effective integration and information exchange between text and images, thereby enhancing the quality and coherence of generated results. In this study's methodology, we delve into the correlation between the cross-attention map and the spatial features of generated images, devising a methodology to accomplish our research objectives.

Chapter 2

Methodology

In this section, we first introduce some preliminary research on text-to-image diffusion models, which involves the principles of diffusion model and the role of cross-attention maps. After that, I depict in detail the implementation of a text-driven solution for accurate image editing, which includes the control of cross-attention in the diffusion model. Finally I demonstrate the feasibility of this method through experimental results and generated images

2.1 The process of stable diffusion model

The diffusion models always contains two key processes which are the diffusion process (forward process) and the denoising process (backward process). In the diffusion process, diffusion models gradually introduces Gaussian noise into the data by a Markov chain for T-step iterations. At the meanwhile, the denoising process constantly generates samples from the Gaussian noise by means of a learnable model. So following the two process, such models can be conditioned on the type of content of the input, e.g., in a text-to-image diffusion model, the model can generate images based on the text input. The training objective of a diffusion model can be reduced to a simple form in this function:

$$L_{\text{simple}} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}||\epsilon - \hat{\epsilon}_\theta(x_t, c, t)||_2^2 \tag{1}$$

Here, x_0 represents the data, and c is the additional condition. t represents the time step in the diffusion progress when $t \in [0, T]$, $x_t = \alpha_t x_0 + \sigma_t \epsilon$ represents the noisy data in t step, α_t and σ_t are the predefined functions of t. Once trained, stable diffusion can iteratively generate images from random noise.

For conditional diffusion models, classifier bootstrapping is a simple and straightforward technique that balances image fidelity and sample diversity by utilising gradients obtained from separately trained classifiers. In order to eliminate the need to train classifiers independently, classifierless bootstrapping is often used as an alternative. At the sampling stage, the predicted noise is computed based on the predictions of the conditional model $\hat{\epsilon}_\theta(x_t, c, t)$ and the unconditional model $\hat{\epsilon}_\theta(x_t, t)$:

$$\hat{\epsilon}_\theta(x_t, c, t) = w\hat{\epsilon}_\theta(x_t, c, t) + (1 - w)\hat{\epsilon}_\theta(x_t, t) \tag{2}$$

In diffusion models, classifierless bootstrapping plays a key role in enhancing the image-text alignment of the generated samples. In this study, I utilize an unconditional image generation model called stable diffusion (SD). Compared to pixel-based diffusion models, text-based SD is more efficient because it is constructed on the latent space of a pre-trained autoencoder model and it uses text prompts as the input. In the training phase of stable diffusion model:

1. SD maps the input images from pixel space to implicit vector space using AutoEncoderKL self-encoder to convert RGB images to implicit vector representations.
2. The FrozenCLIPEmbedder text encoder is used to encode the input text word Prompt to generate the vector representation context.
3. Different intensity noise is applied to the implicit vectors of the input image, and then the noised implicit vectors are inputted into the UNetModel to output the predicted noise, which is compared with the real noise information labels to calculate the KL dispersion loss, and the backpropagation algorithm to update the KL dispersion loss.

2.2 Cross-attention

Ronneberger et al. [25] set the U-shaped network to specify a prediction rule for noise ϵ for the diffusion step t, which consists of a noisy image z_t and a text embedding $\psi(P)$. The interaction of the noisy image z_t and the text embedding $\psi(P)$ suggests that the noise prediction is generated by the fusion of two modalities. These two modalities are fused through the cross-attention layer and they can generate an attention map for each token in the embedded text. This suggests that each token in the input text will have its corresponding attention map, and the research can control the image generation by decomposing or combining the attention maps corresponding to these tokens and controlling their share in the diffusion process for the purpose of text-driven image editing.

In machine learning and deep learning, the Softmax function usually describes how to compute the relevance weights of key and query in the attention mechanism. The spatial feature $\phi(z_t)$ of the noisy image are projected into the query matrix Q, and the text embedding is projected into the key matrix K and the value matrix V by learning the linear projection. Q and K are notated to get the attention map M_t. d is the projection dimension of the key and query and M_{ij} defines the weight of the jth token on pixel i. At the end, the attention map M and value matrix V are multiplied to get the final output, which is the updated spatial feature $\hat{\phi}(z_t)$.

In the stable diffusion model, text features from the CLIP text encoder are inserted into the UNet model by feeding into the cross-attention layers. Given query features Q and text features K, the output of cross-attention maps M and spatial feature $\hat{\phi}(z_t)$ can be calculated by:

$$M = \text{Cross-attention map}(Q, K) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{2}$$

$$\hat{\phi}(z_t) = \text{Spatial Feature}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{3}$$

In the functions, W_q, W_k, W_v are the weight matrices of the linear projection layers, then they are used to calculated $Q = W_q$, $K = c_t W_k$, and $V = c_t W_v$, which are the query, key, and values matrices of the attention maps.

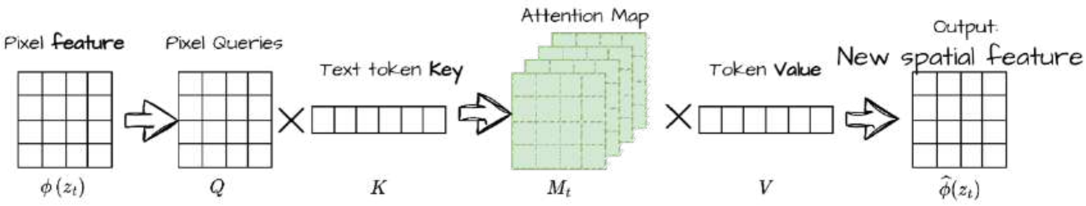


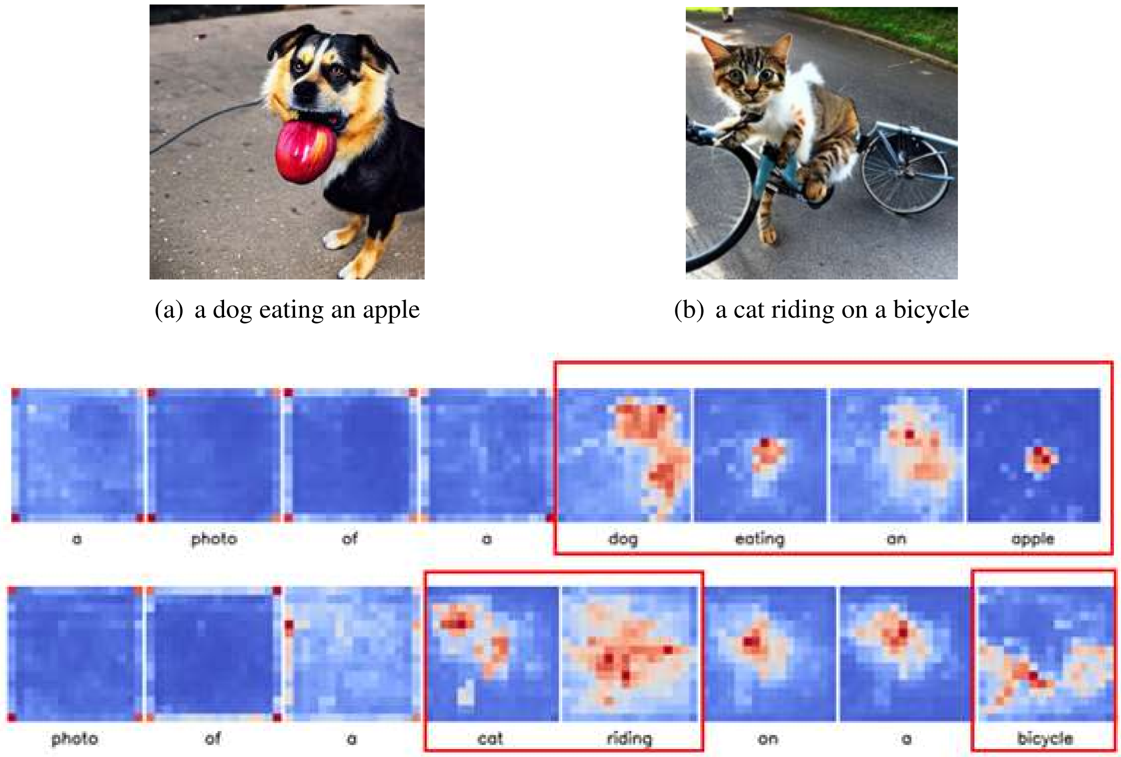
Fig. 2.1: Fusion of text and attention map in diffusion process

To better understand the importance of cross-attention maps in the process of diffusion model, the process can be represented in the Fig.2.1. The spatial layout and shape features of the image, also known as the spatial feature $\phi(z_t)$, are jointly determined by the attention image and the token. And, for the diffusion process of a given text, the text is fixed, that means the token value in the text is also fixed. This indicates that the attention image largely determines the spatial feature $\phi(z_t)$ of the image. When we want to achieve the goal of this study, which is to maintain the spatial layout of the source image in the edited image, this means that the control of the cross-attention map during the process of diffusion becomes the focus of this study. In the next section I will elaborate on the control of the cross-attention maps.

2.3 Controlling of attention maps in diffusion model

2.3.1 Attention map visualization:

In order to better understand the relationship between the attention map and the spatial feature of the image, I tried to visualize the attention maps of the images generated by the SD model, which are shown in Fig.2.2.



(a) a dog eating an apple

(b) a cat riding on a bicycle

(c) Fusion of text and attention map in diffusion process

Fig. 2.2: The images generated by the SD and their attention maps visualization

Fig.2.2, illustrates the interaction between image pixels and text, which consists of images and their corresponding attention maps generated from two sets of text prompts, (a) [a dog eating an apple] and (b) [a cat riding on a bicycle]. I averaged the corresponding cross-attention maps for each token, and in (c), it can be seen that each token in the text prompt has its corresponding cross-attention map, and there is a strong correspondence between them. For example, the attention graphs of the tokens ***dog,eating*** and ***apple*** in (a) have the same shape and structure as the corresponding parts in the generated image. Similarly, the more descriptive tokens in (b) such as ***cat, rating*** and ***bicycle*** also confirm the assumption that the cross-attention graph controls the spatial features of the image. More importantly, in the diffusion process, each token has its own separate cross-attention graph and exists separately and independently from the other tokens, which suggests, to some extent, that this study can consider operations such as substitution, deletion, or addition to the cross-attention graph to achieve the goal of this study, in order to achieve the preservation or modification of the spatial features of the image.

2.3.2 Time step T in diffusion progress

In diffusion models, T usually denotes the diffusion time step. This time step determines the number of steps in the forward and reverse process. In the forward process of the diffusion model, the symbol T denotes the total number of noise addition steps. At each step of the forward process, the system introduces a Gaussian-distributed noise into the image until the image is gradually transformed into an essentially pure Gaussian-distributed noise image. In time step $t \in [0, T]$, the initial stage ($t = 0$) the model acquires the original image by CLIP, and then as the time steps increase, the noisy image obtained at each x_t is obtained by adding a Gaussian-distributed noise to the previous step x_{t-1}.

In the reverse of the diffusion model, in contrast to the forward process, this process denoises a noisy image filled with a Gaussian distribution. $t \in [0, T]$, the initial phase ($t = T$) the neural network acquires the noisy image, and subsequently denoises the noisy image in x_t with each x_{t-1} as the time step decreases. By gradually decreasing T, the increase in noise and the preservation of image details can be balanced during the generation process so that the generated image has both clear details and a natural appearance. The noise image and a simple explanation of the both processes are shown in Fig.2.3.

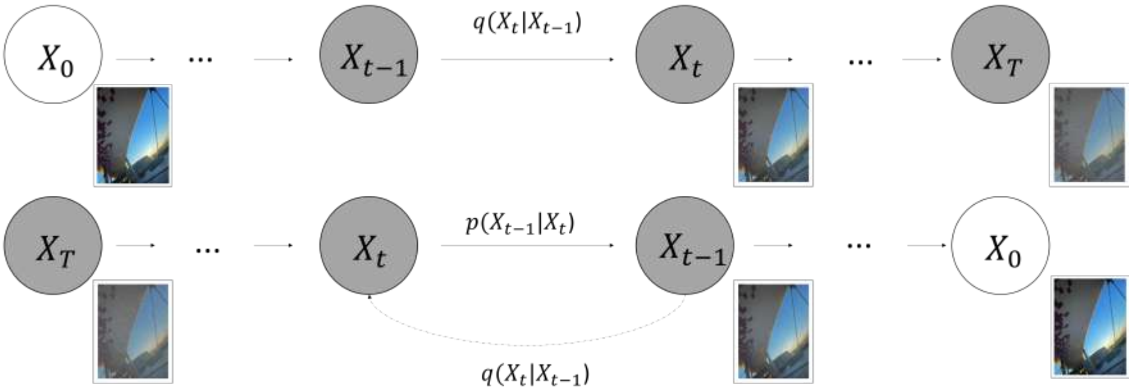


Fig. 2.3: *Top*: forward process; *Bottom*: reverse diffusion process

In the prompt-to-prompt method proposed by Mokady et al. [26], it was found that the structure of the token's cross-attention map was determined at the initial stage in the diffusion process, and the cross-attention map at subsequent stages would only become clearer in terms of its content, while the shape and structure remained unchanged. This suggests that the spatial features of the image are determined due to the fixation of the structure of the cross-attention graph at the early steps, i.e., when time step T's value is large in the diffusion progress. Then this study considers the time steps of the method implementation when performing relevant operations such as substitution or addition to the cross-attention map, controlling the method implementation according to the characteristics of the operation.

2.3.3 Controlling of attention maps

Returning to the observation of the cross-attention map visualization in this study, the cross-attention map determines the spatial feature in the early stage of the diffusion progress. Based on this rule, I consider to control the cross-attention map of the edited text prompt $P*$ during the diffusion process, which involves replacements, additions and scaling of the attention map. This can be simply understood as injecting the attention map M of the original text prompt P into $P*$, such that the edited image $I*$ is generated according to $P*$ while preserving the spatial feature of I. Therefore, this study first attempted to write a general editing framework which satisfies the variability relationship between M and $M*$, I and $I*$ by controlling $M*$ to achieve the objectives of this study. Since this study only considers the use of algorithmic formulations for the control of the attention map and is not prepared to make changes to the noise reduction process of the SD model, this study can focus on the noise addition process.

Let $DM(z_t, P, t)$ be the computation of a single step t of the diffusion process, then the general framework and algorithm can be understood by the follow:

1. Diffusion process: At each t, the original image z_t is noised to produce a sequence of images with varying degrees of noise

2. $DM(z_t, P, t)$: The core computational step of the diffusion process, which accepts as input the current image z_t, text prompt P, time step t and random seed s, and generates the attention image M_t and the noisy image for the next time step $T = t - 1$.

3. $DM(z_t, P, t)\{M \leftarrow M_{ct}\}$: Variant the diffusion process that uses an additional given attention image M_{ct} to overlay the current attention map M. M_{ct} is determined by the $Edit(M_t, M_t^*, t)$ and it depends on the users expected image editing

Algorithm 1 ***Text-driven image editing*** with attention control methods

```
Require: Original text prompt $P$, edited text prompt $P^*$, random seed $s$.
Ensure: Original image $I$ and edited image $I*$.
$z_T \sim \mathcal{N}(0, I)$ unit Gaussian random variable
$z_T^* \leftarrow z_T$
for $t = T, T-1, \dots, 1$ do
    $z_{t-1}, M_t \leftarrow DM(z_t, P, t)$
    $M_t^* \leftarrow DM(z_t^*, P^*, t)$
    $M_{c,t} \leftarrow Edit(M_t, M_t^*, t)$
    $z_{t-1}^* \leftarrow DM(z_t^*, P^*, t)\{M \leftarrow M_{c,t}\}$
end for
return $(z_0, z_0^*)$
```

type. i.e., when we try to edit the text ’dog eating an apple’ to ’cat eating an apple’, we replace the attention map of the dog and cat in the early stage of the edited image diffusion process, the M_{ct} is the attention map of original prompt which contains ’dog’.

4. $P*$, M_t^*: Using edited text prompt M_t^* to generate M_t^*.

5. $Edit(M_t, M_t^*, t)$: Defines this function as a general edit function that accepts as input the attention map of both the original image and the edited image (M_t and M_t^* from step t) and performs some specified form of editing or adjustment on them. The type of edit is determined by text replacement, text addition and image scaling, which will be discussed in next part in detail.

Text replacement:

Text word replacement (change world) belongs to one of the text editing types in the research objectives, which refers to a situation where the user wants to replace a word or phrase in the original text prompt. For example, when the original text prompt P_1 = [dog eating an apple] and P_2 = [cat riding on a bicycle], the user may change the two prompts to P_1^* = [cat eating an apple] and P_2^* = [cat riding on a skateboard]. For P_1 and P_1^*, the user only wants to replace the apple in M_1 generated by P_1 with an orange, while keeping the edited image M_1^* with the same image structure and background as M_1, especially making sure that the shape of dog will not be changed. Similarly, the editing of P_2 follows the same principle.

The main challenge of text word swapping is to generate the content corresponding to P^* while preserving the shape structure of I. To achieve this goal, this study injects the cross-attention map of the original image I into the early diffusion process of P^* such that finally P determines the image spatial features of I^*, P^* determines the image

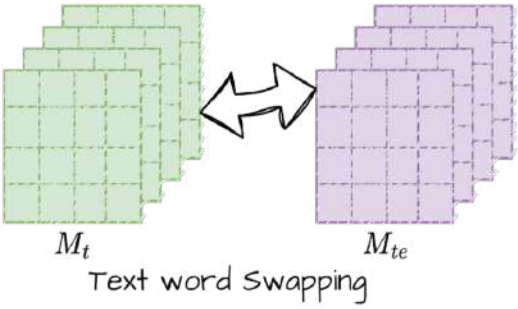


Fig. 2.4: Replacement of cross-attention maps for text before and after editing, inject the green attention map M_t and let it cover the purple attention map belongs to the edited text prompt's tokens.

content of I^*, and the image content is constrained by the spatial features. The formula corresponding to the algorithm can be defined:

$$Edit(M_t, M_t^*, t) = \begin{cases} M_t & \text{if } t > T \\ M_t^* & \text{otherwise} \end{cases}$$

We have already discussed the effect of the time step on the diffusion process in 2.3.2time step T in diffusion progress, and T is decreasing, so we can achieve the goal by adding the time constraints. Additionally, assuming that the number of words the user wants to replace when performing this operation is inconsistent, I would use an alignment function to duplicate and average the attention maps, which would allow a certain degree of aggregate freedom to adapt the new image generation process to the edited prompt, for example, a user editing will replace rabbit with teddy bear.

Text Refinement:

Text refinement refers to the fact that when editing a text prompt, the user wants to enrich the context of the text by adding more word tokens to make the edited image contain more elements or styles. For example, suppose P = [a car on the side of the street], and the user adds some modifiers so that the edit prompt P^* contains a old/Ferrari/wooden car or street in Hawaii/New York/countryside and other similar tokens. The biggest challenge for such an operation is that this study needs to preserve the original structural information of the I generated by P and to ensure that the newly generated I^* contains the content of the image corresponding to the new token in P^*.

To this end, in order to preserve the same details in P and P^*, I inject the attention maps corresponding to the same tokens in P and P^* into the I^* diffusion process. Note that in this part we cannot directly inject the cross-attention graphs of P into the diffusion

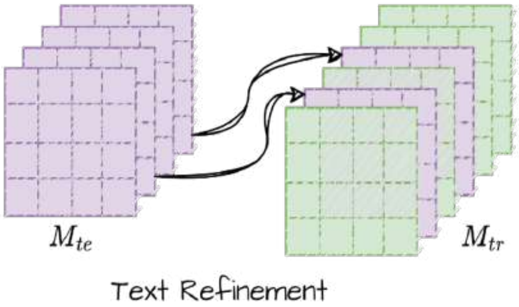


Fig. 2.5: Inject the corresponding tokens' attention maps of the unedited prompts to the edited diffusion process.

process of I^* in order, because we are not sure where the user has inserted a new token in P , and directly injecting all of the attention graphs of P will result in disordered ordering of the attention graphs corresponding to tokens of the diffusion process. Therefore, I use a alignment function $Align_(j)$ accepts the index of the token in P^* and outputs the index of the same token in P.

$$Edit(M_t, M_t^*, t)_{i,j} = \begin{cases} (M_t^*)_{i,j} & \text{if } Align_(j) = None \\ (M_t)_{i,Align_(j)} & \text{otherwise} \end{cases}$$

In $Align_(j)$, j corresponds to the text token and i is an index corresponds to a pixel value. Similar to the ***Text replacement*** method, we can control the number of injection steps, i.e. the time step t, to achieve a variety of image editing according to different word refinement characteristics, such as adding new objects to an image or changing the style of a picture from oil painting to sketch, etc. time step t to achieve a variety of image editing based on different word refinement features, such as adding new objects to an image or changing the style of an image from an oil painting to a sketch, and so on.

Text semantic re-weighting:

Text semantic effect modification refers to the fact that for a textual cue P, the user may want to enhance or diminish the emotional semantic effect of the token in P, which will result in increasing or decreasing the image's sensitivity to the token to make the image have different effects. For example, when our textual cue is a smiling bear, we want to make the bear smile very happily, so we need to control the attention graph of smiling.

To achieve this goal, we can control the size of the structure of the attention map to make the emotional semantics of the token larger or smaller. This approach can be done by setting a control parameter k to scale the size of the attention graph of token j. Thus, the

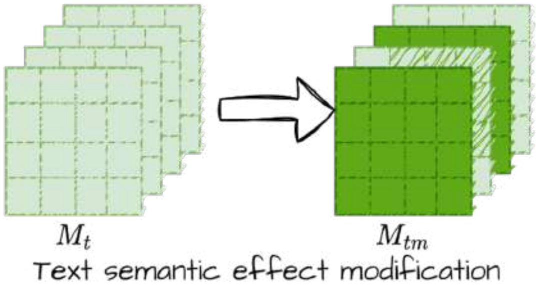


Fig. 2.6: Scale the edited token's attention map to re-weight the semantic effect of the edited token.

editing formula can be written as

$$Edit(M_t, M_t^*, t)_{i,j} = k \times (M_t^*)_{i,j}$$

2.3.4 Self-attention injection

This study focuses on controlling the cross-attention map to manipulate the spatial features of images, achieving various editing effects through different methods as outlined earlier. However, typical diffusion models incorporate both cross-attention and self-attention layers. The self-attention mechanism allows the model to prioritize different regions within the input image, guiding the image generation process towards producing more realistic outputs. By understanding the relationships between different regions, the model can enhance details such as object boundaries and textures, leading to higher-quality images.

Our study addresses image structure preservation by adjusting the cross-attention layer during the early stages of the diffusion process. Additionally, to better preserve the original content while adapting to newly generated features, we explore injecting self-attention corresponding to the initial cue into the diffusion process of the edited image. By varying the degree of self-attention injection from 0 to 100%, we observe enhanced naturalness, quality, and realism in the generated images, particularly evident at a time step of 25%.

Chapter 3

Experimental results

3.1 Experiment environment

In this study, we utilized the Stable Diffusion v1-4[1] model as the experimental foundation, with a diffusion process comprising 50 steps. Experimentation was conducted within the Jupyter Notebook, using Python version 3.11.4. The necessary Python packages, including torch, diffusers 0.14.0, opencv-python, transformers, and ipywidgets, were employed during the experimental process. Our experimental setup consisted of an Intel(R) Core(TM) i9-10929X CPU @ 3.50GHz as the CPU and an NVIDIA GeForce RTX 3090 as the GPU.

3.2 Results

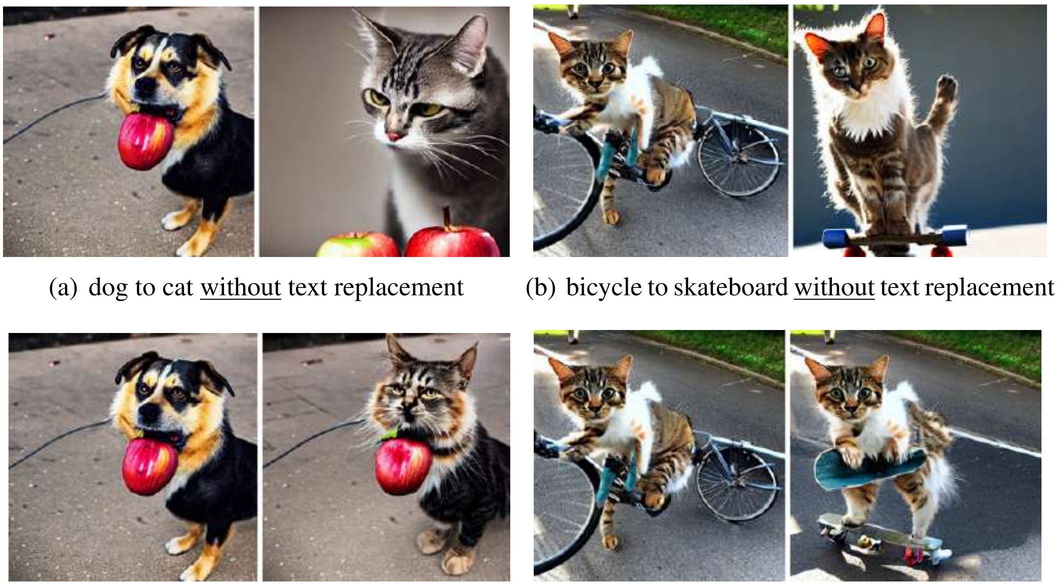
In this section, this study experiments with the three methods mentioned in the methodology using the stable diffusion model to demonstrate their practical application.

3.2.1 Text replacement

The text replacement method provides the user with the opportunity to make replacements for object elements in an image. We set two sets of text prompts to demonstrate the feasibility of text replacement. Suppose the text prompt p_1 = [dog eating an apple], p_2 = [cat riding on a bicycle]. In p_1 we replace dog with cat and in p_2 we replace bicycle with skateboard. In Fig.3.1, the top two figures are a comparison of the images before and after editing without text replacement, while the bottom two figures are a comparison

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

of the images before and after editing with text replacement. It is observed that the this method allows the user to replace words while preserving the content, background and spatial features of the source image.


(a) dog to cat without text replacement

(b) bicycle to skateboard without text replacement

(c) dog to cat with text replacement

(d) bicycle to skateboard with text replacement

Fig. 3.1: *Top*: After editing the text prompts, the edited image is generated directly by stable diffusion, the edited image only meets the text requirements but does not maintain the structure of the image before editing. *Bottom*: The edited image is generated after injecting the cross-attention map of the original text into the diffusion process, which maintains the structure and scene information of the image before editing, it conforms to the research objective

Applications in different scenarios:

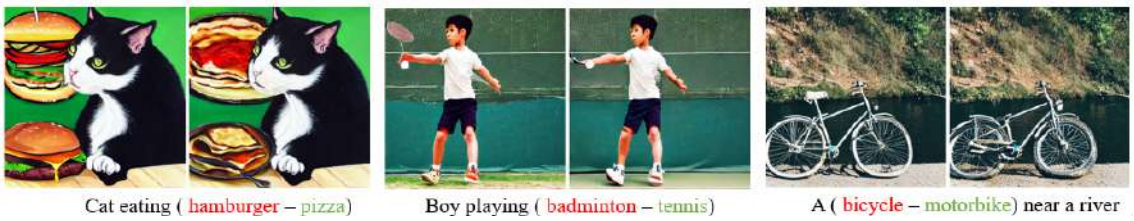

Cat eating (hamburger – pizza)
Boy playing (badminton – tennis)
A (bicycle – motorbike) near a river

Fig. 3.2: Different styles' images with text replacement method

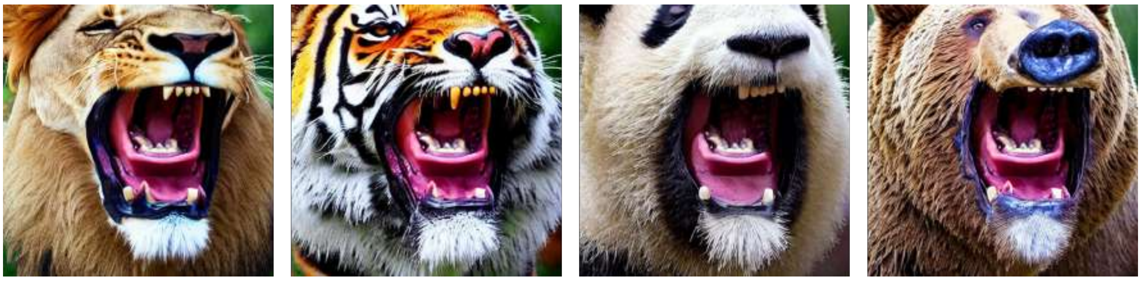


Fig. 3.3: 1. source prompt = "The photo of a lion laughing with a big mouth", edit to 2. tiger; 3. panda; 4. bear

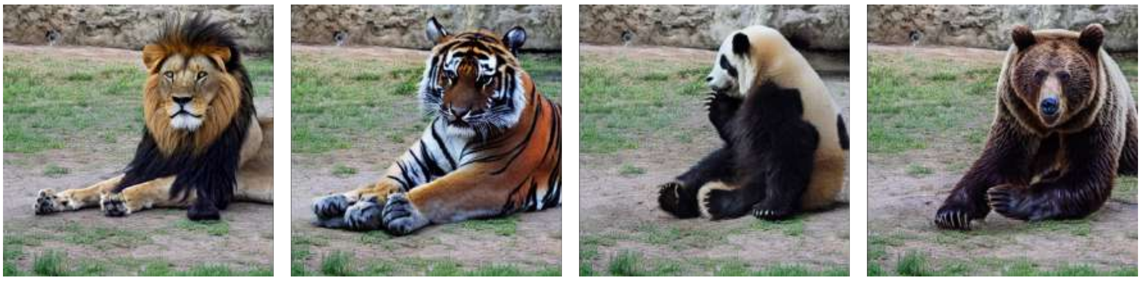


Fig. 3.4: 1. source prompt = "The photo of a lion in a zoo", edit to 2. tiger; 3. panda; 4. bear

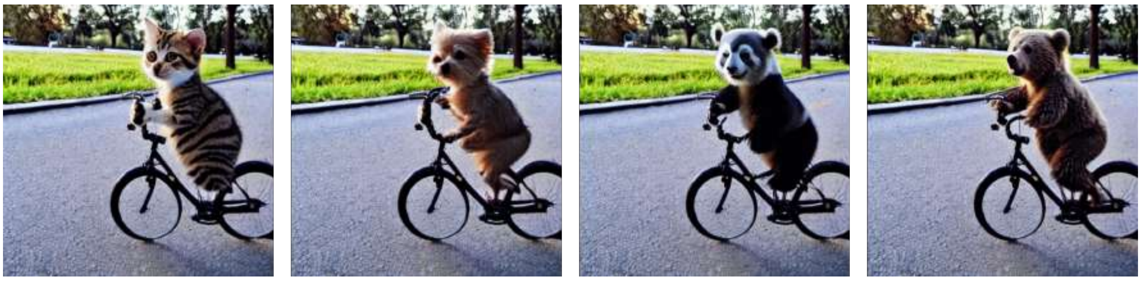
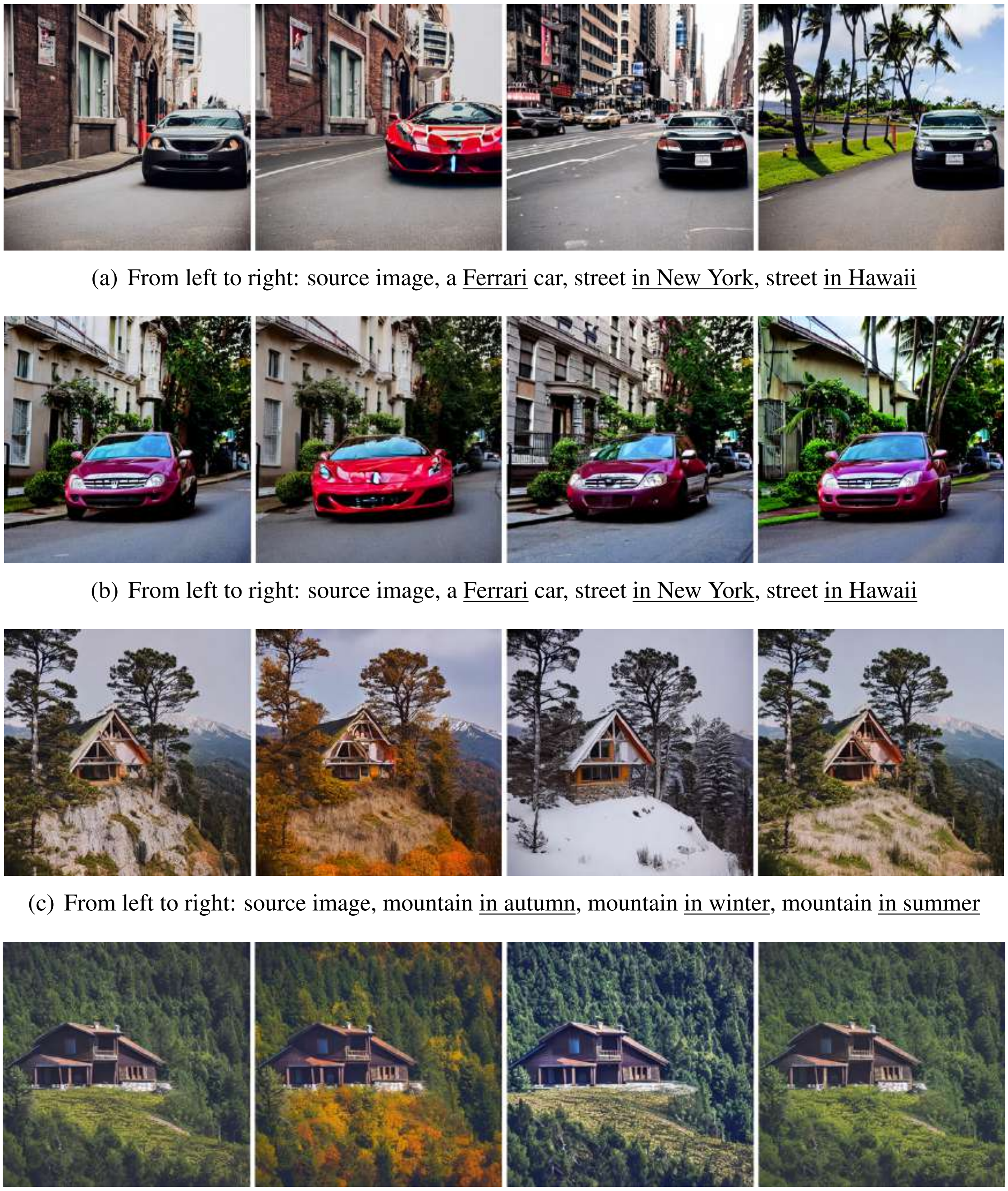


Fig. 3.5: 1. source prompt = "The photo of a small cat riding on a bicycle", edit to 2. dog; 3. panda; 4. bear

3.2.2 Text refinement

The text refinement method provides users with the ability to add words or phrases to the text prompt to edit the image to give it more contents and structure. We set text prompt p_1 = [a photo of a house on a mountain], p_2 = [a car on the side of the street]. We use two generation styles to generate images for p_1 and p_2, and then we add phrases to p_1 and p_2. In Fig.3.6, for p_1, we added seasonal modifiers to change the climatic context of the image. For p_2, we added car brands and street locations to enrich the features of the image. We used different random seeds to set different image styles and get the diverse results.



(a) From left to right: source image, a Ferrari car, street in New York, street in Hawaii

(b) From left to right: source image, a Ferrari car, street in New York, street in Hawaii

(c) From left to right: source image, mountain in autumn, mountain in winter, mountain in summer

(d) From left to right: source image, mountain in autumn, mountain in winter, mountain in summer

Fig. 3.6: Different text refinements ways with Different diffusion styles. *Top 2 figures*: The text edit adds constraints to either the car or the location, but the edits produce images that maintain the original graphic structure and try to avoid image modifications to the roads and original buildings within the image. *Bottom 2 figures*: The editing of these two sets of images is much clearer, and with the addition of the seasonal conditions, stable diffusion maps the seasonal changes by changing the colours of the trees and mountains.

Applications in different scenarios:

Global editing: style change

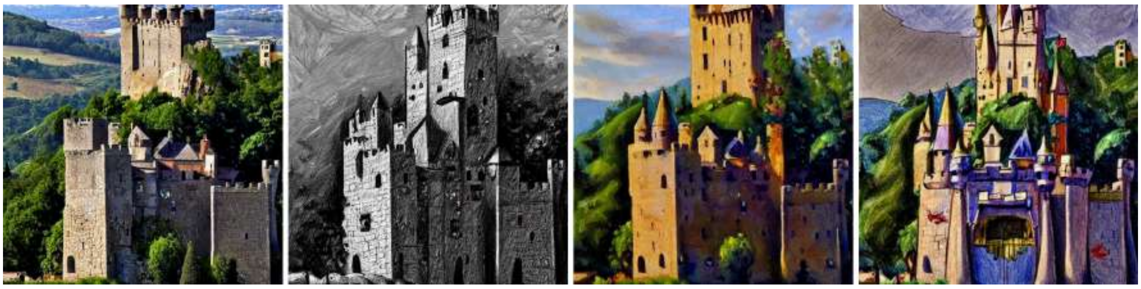


Fig. 3.7: 1. source prompt = "a medieval castle", edit to 2. "a sketch picture of ..."; 3. "an oil painting of ..."; 4."Children drawing a disney style of ..."

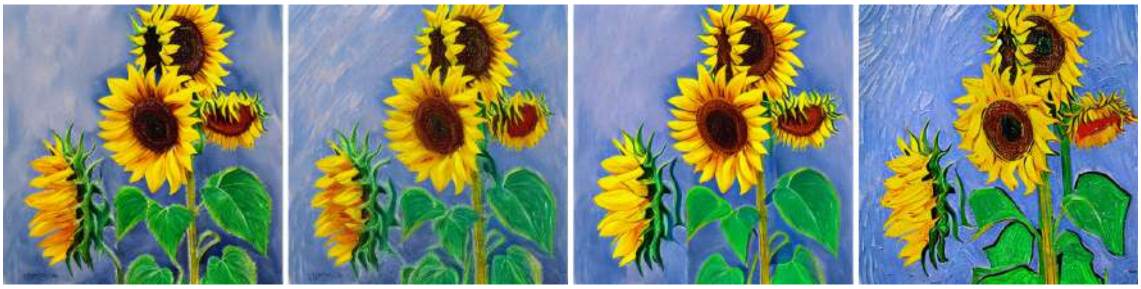


Fig. 3.8: 1. source prompt = "sunflowers", edit to 2. "a picture of sunflowers in the style of Monet"; 3. "a picture of sunflowers in the style of Picasso"; 4. "a picture of sunflowers in the style of Van Gogh"

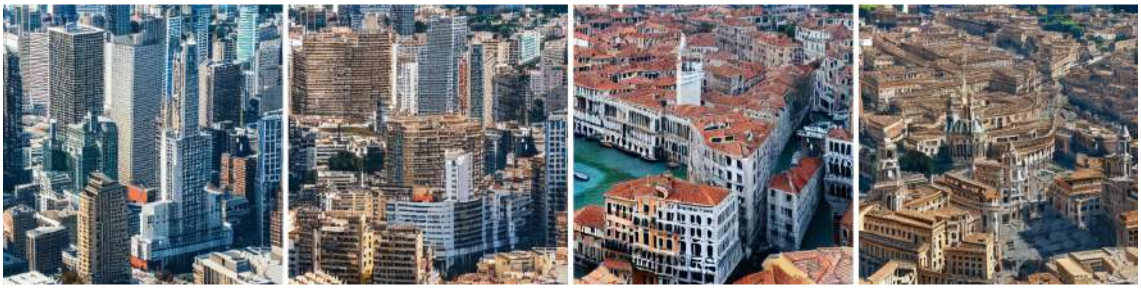


Fig. 3.9: 1. source prompt = "photo of lots of high buildings" edit to 2. "photo of lots of high buildings in Spain"; 3. "photo of lots of high buildings in Venice"; 4. "photo of lots of high buildings in Vatican"

Local editing: elements and contents addition:

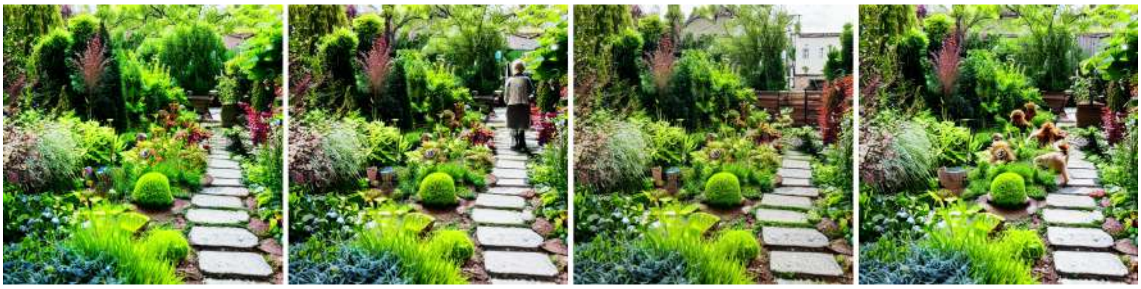


Fig. 3.10: 1. source prompt = "photo of a small garden" edit to 2. "photo of a small garden with a old lady walking in it"; 3. "photo of a small garden with a house in the background"; 4. "photo of a small garden with dogs and cats"

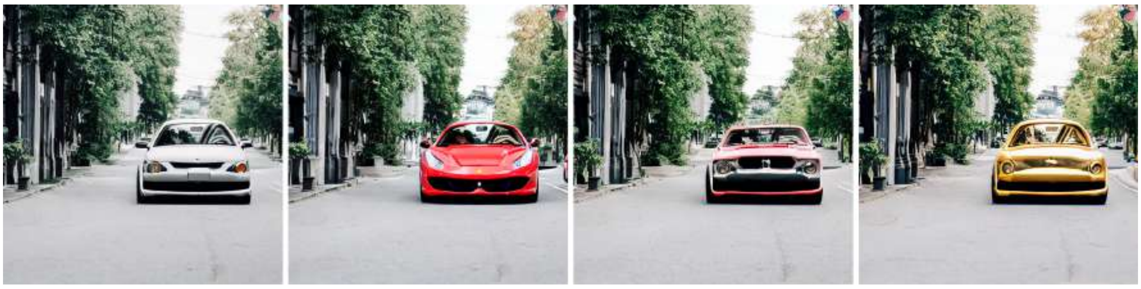


Fig. 3.11: 1. source prompt = "a car on the side of the street" edit to 2. "a Ferrari car"; 3. "a muscle car"; 4. "a golden car"

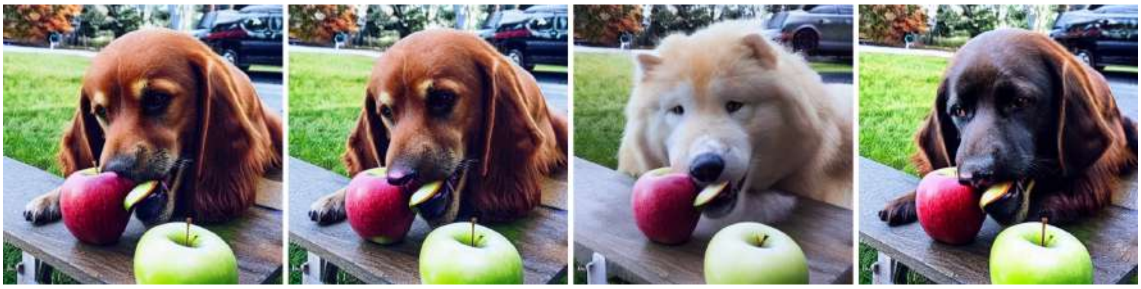


Fig. 3.12: 1. source prompt = "A photo of a dog eating an apple" edit to 2. "a Laburador dog"; 3. "a Samoyed dog"; 4. "a black dog"

3.2.3 Text semantic re-weighting

The text semantic re-weighting method strengthen or weaken the semantic effect of elements in an image by scaling the corresponding cross-attention map. For example, when we write text prompt p_1=[a smiling teddy bear], p_2=[a lush tree]. We can scale the attention maps for the words smiling and lush when we want to make the teddy bear smile more obvious and the tree's branches more lush. First, we observe the result of this idea.

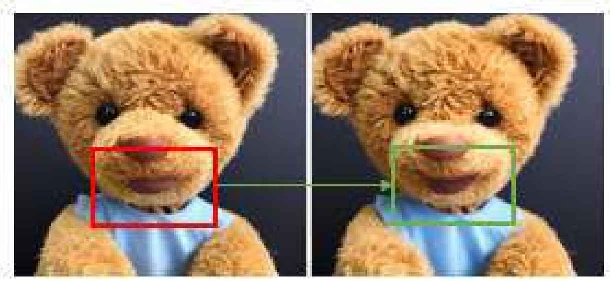


Fig. 3.13: left to right: strengthen the smiling semantic

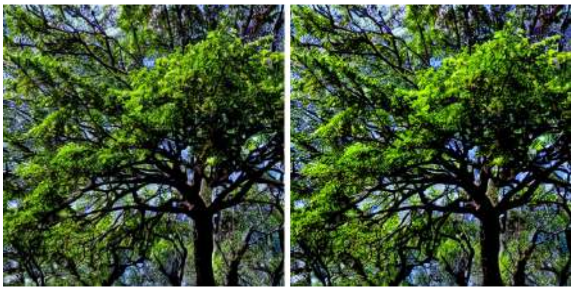


Fig. 3.14: left to right: strengthen the lush semantic

In the observations of Fig.3.13 and Fig.3.14, we can see that the bear's smile has become wider, but the tree has not changed significantly. This is because in the diffusion process of the stable diffusion model, for p_1, smiling is materialised though smiling is usually an adjective in our understanding, it was treated as a noun in the diffusion process, and the attention map enlarged the mouth region corresponding to 'smiling', thus producing better results. However, for p_2, lush as an adjective does not accurately visualise all the leaves in the tree, and therefore scaling the attention map corresponding to lush did not yield satisfactory results. Therefore, we try to change p_2 to p_2^* = [a photo of a tree full of leaves] and try to scale leaves to see if semantic enhancement can be achieved. On the other hand, we modify p_1 to p_1^* = [a boy playing happily with his dog] to observe the situation when adjectives happily cannot be accurately visualised as nouns, which cannot be directly achieved by the text semantic re-weighting method.

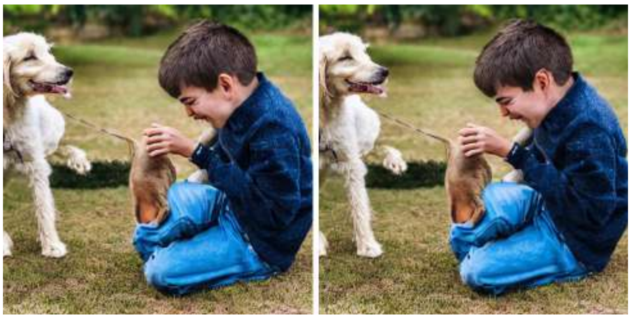


Fig. 3.15: left to right: strengthen the happily semantic

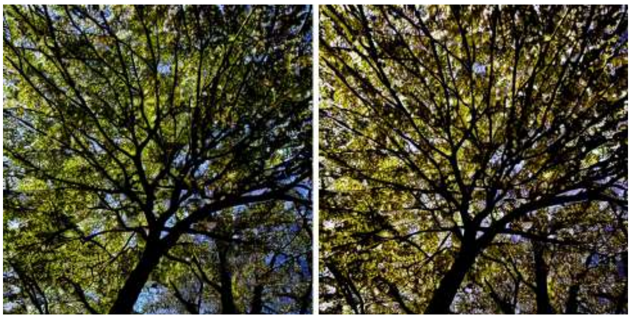


Fig. 3.16: left to right: strengthen the leaves semantic

Fig.3.15 and Fig.3.16 confirm our earlier conjecture that the semantic effect of the image changes in the opposite direction when happily can no longer represent a smile and leaves can represent the lushness of the tree. This suggests that when we try to use the text semantic re-weighting method, we would be better off choosing tokens that can be figuratively edited, which helps us to better control the attention map of the edited text to achieve research objectives.

In a conclusion, we present the research performance through a large number of experimental results on image editing. In these results, the text replacement and text refinement methods perform well in various scenarios and do not require much text editing by the user. However, for the text semantic re-weighting method, we need to pay attention to whether the words have a clear attention graph counterpart in the diffusion process. Especially when editing adjectives, this method does not perform well.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

In this work, the research strives to achieving precise image editing solely driven by text-based inputs. To this end, we conducted an in-depth analysis of the relationship between the cross-attention layer and spatial features within the text-to-image diffusion model. Within the diffusion model, the cross-attention map governs the spatial features of the image during the early stages of the diffusion process, while the self-attention map predominantly influences the image's content generation. Building upon this observation, we introduced three methods for controlling the cross-attention map, demonstrating how to control the spatial structure of generated images during the diffusion process. Leveraging the method algorithmic framework, users can obtain image editing results that maintain the original spatial structure of the original image solely through editing text prompts. These editing capabilities encompass: 1. Image elements' replacement achieved through text replacement, 2. Content augmentation and style modification achieved through text refinement, and 3. Senmantic modification of image elements through text semantic re-weighting.

Through extensive experimentation encompassing both local and global edits, as well as adjustments to image content and style, the method has showcased successful multidimensional editing effects across various experimental scenarios. This proves the method is capable of achieving accurate image editing driven only by text-based editing. However, I acknowledge the existence of limitations across different dimensions, the strengths and weaknesses of this method can be summarized as follows:
Strengths:

1. This method exclusively requires control over the cross-attention maps within the

forward process of the diffusion model through code implementation, eliminating the necessity for model retraining during the denoising process. This circumvents the time-consuming feature extraction and model training inherent in traditional image editing methods, resulting in a substantial enhancement in editing speed.

2. This method preserves the content information of the original image while simultaneously governing the spatial features of the edited image. By doing so, it circumvents the risk of losing original content, which often occurs with the removal of structural information in mask-based methods.

Weaknesses:

1. For the text replacement method, when there are significant differences in size between the elements the user intends to replace (e.g., apple and watermelon), the edited image may not meet the user's expectations. Despite the watermelon appearing in the position of the original apple in the image, its size will remain the same as that of the apple. This implies that when the type and size differences between the elements edited by the user are substantial, the generated image may maintain its structure but may not accurately reflect the user's intent.

2. In the case of the text semantic re-weighting method, changes in the expression of emotional effects in the image corresponding to certain words may not be significant before and after editing. This is because the tokens the user intends to edit are predominantly adjectives. However, adjectives' cross-attention maps do not exhibit a fixed structure like objects, so scaling them may not yield optimal results.

3. Currently, the method yields different outcomes when faced with different random seeds in the diffusion model. Different random seeds generate images with different styles. As a result, the method may encounter distortions in some cases, which are related to the training dataset and process of the diffusion model itself.

4.2 Future Work

Based on the summary of the strengths and weaknesses of the research in the previous subsection and the development of the text-to-image diffusion model, I believe in the future, this research can continue in the following directions:

- **Construct a labeled dataset**: Suppose we aim to deploy this method online to provide users with an intuitive and convenient image editing functionality. To meet diverse editing needs and mitigate the issues mentioned in drawback 1, we plan to build a labeled dataset. This dataset will encompass tokens that users may edit, categorized based on their characteristics. For instance, we can categorize tangible objects by size, where strawberries and cherries belong to small fruits, while watermelons and durians belong to large fruits. Thus, when users replace tokens between different size categories of fruits, our code can automatically recognize the size differences among fruits and adjust the size of attention maps accordingly, ensuring that the edited image's fruits conform to users' expectations in both structure and size.

- **Improve editing accuracy and realism**: Current methods sometimes lead to distortions in the generated images, and elements in the edited image are limited by the structural shape of the original image, thus requiring highly accurate text prompts from the user. To solve this problem, we can consider introducing adaptors designed in the DreamBooth and Textual Inversion models into the diffusion process. This helps to generate content that is more realistic and more in keeping with the theme of the original image.

- **Increase the range of applications of the method in diffusion models**: This study is conducted based on the stable diffusion model, and future research could consider extending the method to more advanced diffusion models. this would involve optimisation of the methodological parameters as well as further investigation of the attention layer.

Reference

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo, “Raphael: Text-to-image generation via large mixture of diffusion paths,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv preprint arXiv:2112.10741*, 2021.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents. arxiv 2022,” *arXiv preprint arXiv:2204.06125*, 2022.

[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[6] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, and B. K. Ayan, “Parti: Pathways autoregressive text-to-image model.”

[7] O. Avrahami, O. Fried, and D. Lischinski, “Blended latent diffusion,” *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–11, 2023.

[8] A. Photoshop, A. Premiere, C. Pro, and A. B. CS, “Adobe®,” *Praha: Adobe, c2019 [cit. 2019-05-21]. Dostupné z: https://www. adobe. com/cz/products/photoshop. html*, 2022.

[9] I. M. Howat, A. Negrete, and B. E. Smith, “The greenland ice mapping project (gimp) land classification and surface elevation data sets,” *The Cryosphere*, vol. 8, no. 4, pp. 1509–1518, 2014.

[10] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[11] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[15] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," in *European Conference on Computer Vision*. Springer, 2022, pp. 88–105.

[16] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[20] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *arXiv preprint arXiv:2210.11427*, 2022.

[21] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework.” *ICLR (Poster)*, vol. 3, 2017.

[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[23] S. Targ, D. Almeida, and K. Lyman, “Resnet in resnet: Generalizing residual architectures,” *arXiv preprint arXiv:1603.08029*, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[25] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18.* Springer, 2015, pp. 234–241.

[26] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross attention control,” *arXiv preprint arXiv:2208.01626*, 2022.