Master's Thesis

Sketch-Based Scene Image Generation with Two-Stage Latent Diffusion Model

ZHANG TIANYU

Supervisor HAORAN XIE

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

August, 2023

Abstract

Recent text-to-image models can generate high-quality, diverse result images only based on text prompts. However, it is difficult to correctly interpret instructions specifying the layout of compositional scene images using only text. When describing individual objects, text often requires lengthy sentences to convey constraints effectively. However, even lengthy text descriptions for complex scene images still fail to enable the diffusion models to comprehend the intricate spatial relationships and relative scales among objects. In addition, there is the issue of object loss in the conventional text-to-image diffusion models.

In this thesis, we propose a sketch-based method to control the position of corresponding objects in image generation and solve the issue of object disappearance in state-of-the-art diffusion models. The pre-trained text-to-image latent diffusion model was utilized by the proposed method as the image generator without additional fine-tuning or training. We manipulate the cross-attention layers used by the model to connect textual and visual information via the input sketch and guide the image reconstruction with the given desired layout. Specifically, we partition the model into two stages. In the feature extraction stage, the sketches are segmented into individual objects using the image segmentation approach, and the obtained bounding boxes and labels are then used as position-guided inputs to the attention layers of the diffusion models. In the image generation stage, the proposed model utilizes Latent Diffusion Model (LDM) as the generator to generate corresponding images. In the early stages of the diffusion process, the object's position is influenced by the sketch's object position guidance in generating the attention maps.

We conduct experiments quantitatively and qualitatively to evaluate the proposed model's effectiveness. In qualitative experiments, we show the position control of the proposed method. Compared with LDM, the proposed model can control the precise positions of objects in the generated image and effectively solve the issue of object loss. We also explore that the positions of objects in the generated image are determined early in the diffusion process. In addition, We further demonstrate its versatility by changing the position relationships and relative scales in sketches. We demonstrated that after altering the spatial relationships and relative scales among sketch objects, the generated objects in the image could change accordingly, even if it contradicts the content of the text. In quantitative experiments, we compared the proposed model with LDM and Generative Adversarial Networks (GANs).

We show excellent image quality (21.04 FID value) than GANs (143.1 for pix2pix, 141.5 for SketchyGAN, 87.6 for SketchyCOCO).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image generation is currently in a stage of rapid development with new methods constantly emerging. Researchers aim to improve generated images' quality, diversity, and controllability, and explore broader application areas[10, 11, 12], such as face generation[13, 14], image restoration[15], and image style transfer[16]. In the early stages of image generation, the methods relied on handcrafted feature extraction and statistical models. For instance, the textures of input images were modeled and used to synthesize new images with similar textures[17]. The development of deep learning-based approaches, particularly Variational Autoencoders (VAE), autoregressive models, and Generative Adversarial Networks (GAN), has advanced and improved image generation approaches. VAEs stood among the pioneering deep learning-based generative models. VAEs combine the encoder-decoder architecture with probabilistic graphical models to learn latent representations for generating images. VAEs can produce relatively realistic images but suffer from issues like blurry images and latent variable collapse[18, 19, 20]. Autoregressive models are sequence generation models where each element in the generated sequence depends on the previously generated elements. The latest autoregressive Transformer model, CogView, has been introduced as a solution for the task of generating images from text[21]. CogView demonstrates excellent performance in text-image ranking, style transfer, and super-resolution. One drawback of CogView is the sluggish generation process, which is typical for auto-regressive models due to their token-by-token image generation approach. GANs marked a milestone in the field of image generation. GANs consist of a generator and a discriminator, which engage in an adversarial training process to make the generator generate images that progressively resemble real data distribution[22, 23]. In addition, conditional generative models allow additional conditions to be specified during image generation to increase the control and flexibility of the generation process,

such as providing sketches or text descriptions to control the features of generated images[24, 25].

The diffusion model is undoubtedly one of the most revolutionary technologies that have surfaced in the past few years. In computer vision and computer graphics, new text-to-image generation methods have demonstrated remarkable image quality. Such as Stable Diffusion (SD)[1], DALL-E[26], and GLIDE[27], that are shown in Figure 1.1 and Table 1.1. In particular, we report both SD and SD-2.1, as SD-2.1 enhances negative prompts and portrait accuracy, making the images more refined and allowing non-standard resolutions. Inspired by non-equilibrium thermodynamics, Sohl-Dickstein et al. defined a Markov diffusion step chain, which introduces random noise to the data samples gradually and learns the inverse diffusion process to reconstruct the desired image samples from the noise subsequently[28]. Denoising Diffusion Probabilistic Models (DDPM) improves the quality of generated images further by introducing denoising priors on diffusion models[29]. However, generating samples from DDPM by following the reverse diffusion process is prohibitively slow. Denoising Diffusion Implicit Models (DDIM) introduce invertible mapping to enhance both the quality and efficiency of image generation[30]. In addition, Latent Diffusion Model operates the diffusion process in the latent space instead of the pixel space, leading to lower time costs and faster inference process[1]. Presently, diffusion models have surfaced as the latest cutting-edge category of deep generative models. These models have disrupted the long-standing dominance of GANs in the demanding field of image synthesis and have demonstrated promise across various domains, such as computer vision[31], multi-modal modeling[32], and natural language processing[33]. In addition, diffusion models have the capacity to greatly amplify the productivity of professional artists and have attracted widespread interest from the general public in practical applications such as art design and creation.

Scene image generation is a challenging task in the field of image generation. Scene images typically involve the combination of multiple objects and elements, such as humans, animals, and backgrounds, with contextual relationships between them. In scene image generation, the model needs to capture various elements and details in the scene while maintaining the realism of the generated image[7]. Additionally, the model must understand and preserve semantic connections between objects during the generation process, ensuring overall coherence and consistency in the generated images. To address these challenges, researchers have proposed innovative methods, such as attention mechanisms[34], multimodal information fusion[35], and graph neural networks[36], to enhance the performance and quality of scene image generation models.

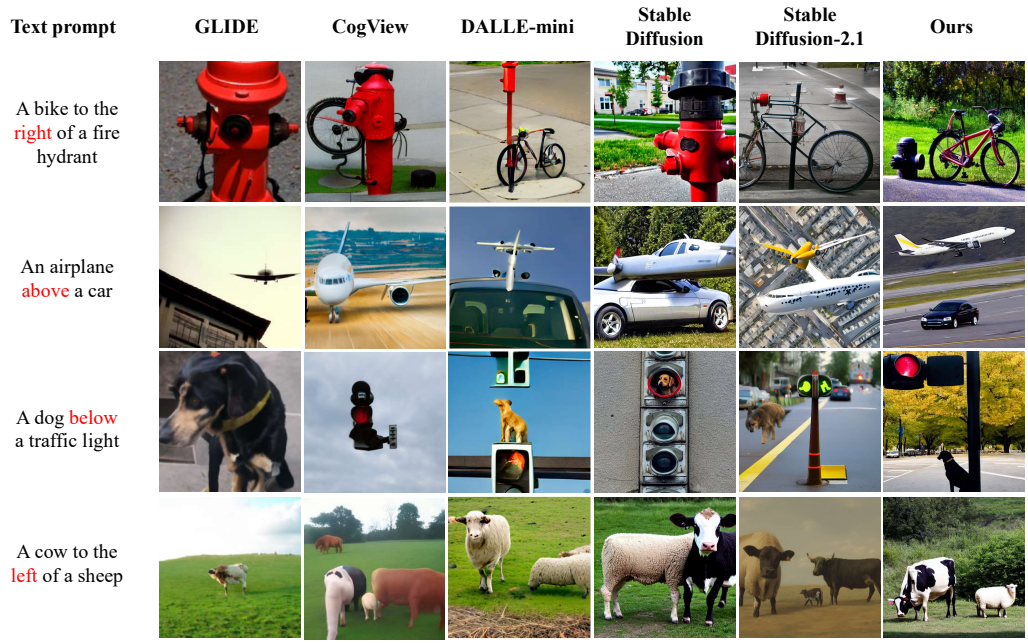


Figure 1.1: The generated results of typical text-to-image generative models. Most of the text-to-image models can not comprehend the corresponding position relationships in the text prompt. However, we finish position guidance for the objects of generated images by the cross-attention maps.

Model	FID(↓)
GLIDE[27]	12.24
CogView[21]	27.10
DALLE[26]	17.89
Stable Diffusion[1]	12.63

Table 1.1: The FID values of well-known text-to-image generative models evaluated on the MS-COCO dataset.[9].

Figure 1.2: The generated results of Stable Diffusion V2.1[1]. As the textual prompts become increasingly complex, Stable Diffusion struggles to comprehend the depicted scene images well and encounters the issue of object omission (the red part is newly added text prompts).

Despite the successes, the powerful pre-trained diffusion models still lack a high level of control that can guide the spatial properties of the scene images. Specifications in text-based image generators of diffusion models are textual. Converting text accurately into visual content is a complex task. Models need to precisely understand the semantics and context of the text, and map it to appropriate image feature representations. While the text is relatively easy to obtain and possesses a vast library of high-level concepts, the text is not an excellent way to express fine-grained visual details in images. As shown in Figure 1.2, scene images typically have multiple objects, backgrounds, and environmental elements, exhibiting high complexity and rich semantic information. For complex scenes, lengthy and intricate text descriptions are often required, involving complex semantic relationships and multiple objects[37]. Generating models struggle to maintain consistency and coherence when faced with long textual descriptions, resulting in issues of blurry or inaccurate generated results and object loss.

In fact, in Stable Diffusion[1], current state-of-the-art image generators struggle to effectively comprehend straightforward layout instructions specified in text form. As an illustration, when given the text prompt "a cow is grazing on the left of the dog," there is indeterminately about whether the resulting positional arrangement of the objects will align with the intended layout. As shown in Figure 1.3, the position between objects is changed from "left" in the text to "up" in the image. In the case where text is difficult to fully control the positional guidance, the typical text-to-image generative model lacks the control method in the generation process. This is mainly because diffusion models lack explicit positional guidance during the image generation process. Diffusion models belong to the category of probabilistic generative models, where the core idea is to iteratively generate real images from noisy images. At each step, the model focuses on updating the image's pixel values without considering the pixels' positional information.

As mentioned above, the current diffusion models face the following issues: 1) the text prompts are difficult to describe the semantic information, especially in scene images; 2) text-to-image generation models lack position control of generated results; 3) diffusion models may lose the objects that depicted in text prompts. To solve these issues, we employed multi-modal and two-stage image conditioning methods. Research on multi-modal image generation involves integrating information from multiple perceptual modalities into image generation tasks. These perceptual modalities can be various forms of data, including images, text, and sketches, providing complementary information to help generate richer and more realistic images[38]. Multi-modal methods are widely applied in conditional image generation to enhance the image quality and effectiveness of image synthesis[39]. Sketches,

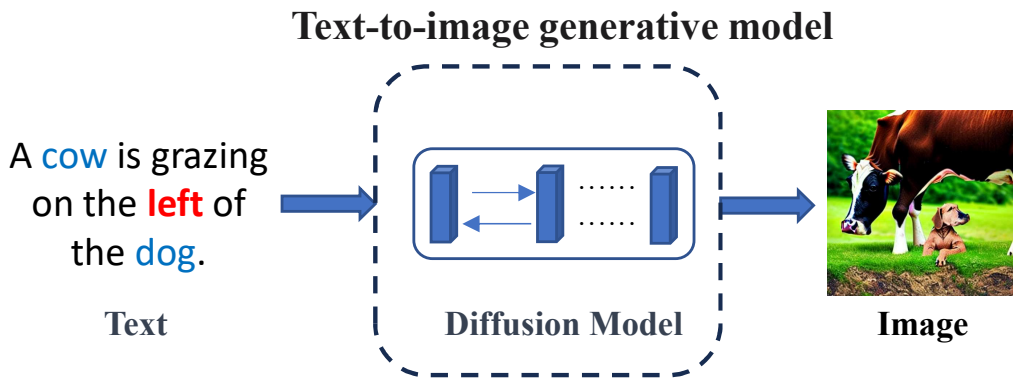


Figure 1.3: A typical text-to-image diffusion model[1] (existing models have poor control over details).

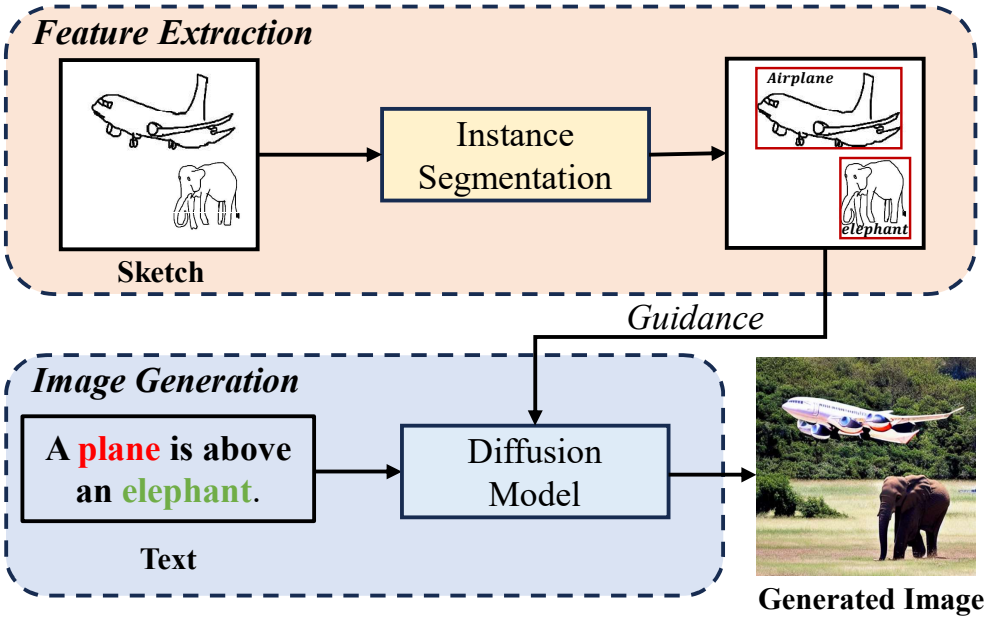


Figure 1.4: Based on the diffusion model, the proposed method is guided by the sketch's segmentation. The proposed method does not necessitate any further training of the pre-trained text-to-image diffusion model.

as an intuitive way of human-computer interaction, express the composition of scenes and relationships between objects. Compared to other inputs, sketches more directly capture users' ideas. Sketches can supplement information that is difficult to describe in text and provide precise location guidance for diffusion models. In addition, the two-stage model decomposes complex image generation tasks into two relatively simple stages, making the model easier to adjust and optimize. In the first stage, it can learn an effective representation of the conditional information, which better guides the generation of images in the second stage[40, 25].

In this thesis, we propose the sketch-based scene image generation method with two-stage latent diffusion model. As shown in Figure 1.4, we try to intervene in the image generation process by adding sketches as new control conditions and altering the attention layers in the diffusion process. In the first stage, we utilize instance segmentation to extract object locations and labels from sketches and encode them as position guidance of the generation process. In the second stage, the pre-trained LDM generates images according to the input text prompts, where the objects' positions will follow the position guidance of the sketches. Our proposed method gets reliable layout control without the need for additional training, while still maintaining the quality of the generated images. The advantage of our method is that it provides the ability to precisely control the location of output images via input sketches, making the generated results more in line with requirements.

We list our main contributions as follows:

- We propose a sketch-based scene image generation model, which intervenes with the spatial properties in attention layers of diffusion models to control the generated objects' positions.
- The proposed model can effectively improve the object loss issue that occurs in the diffusion models.
- We conduct both quantitative and qualitative evaluations to assess the effectiveness of our approach in achieving image position control..

Chapter 2

Related Works

2.1 Diffusion Model

In image generation, diffusion models outperform GANs and VAEs. It is effective and easy to implement, producing images of excellent quality. First introduced by Sohl-Dickstein et al.[28] and later advanced by Song et al.[41] and Ho et al.[29]. In recent times, numerous text-image models of significant scale have surfaced, such as Imagen[42] and DALLE-2[43], demonstrating unprecedented semantic generation.

Diffusion models attempt to transform a simple distribution of random noise into data samples through a series of transformations. It mainly consists of two processes: forward diffusion process and reverse denoising process (inference process). In forward diffusion process, a random image is sampled from the data distribution, and Gaussian random noise is gradually added to the image through a fixed process until it becomes pure noise. In reverse denoising, starting from pure noise, the process gradually restores it to a real image. As shown in Figure2.1, diffusion models[44] iteratively blurs and adds noise to the image, then gradually restores the details of the image to generate the final image samples. Specifically, diffusion models simulate the diffusion process by introducing a time step. The generation process starts with a random noisy image, and the image is updated based on the current image state and the known noise at every time step. The diffusion model gradually restores the image to its original state through multiple iterations and gradually reduces the noise intensity.

Diffusion models can well preserve the texture and details of the image, and the generated results have good visual effects. These models were also able to produce increasingly diverse images and were shown to be immune to mode collapse. However, diffusion models also have some limitations. First,

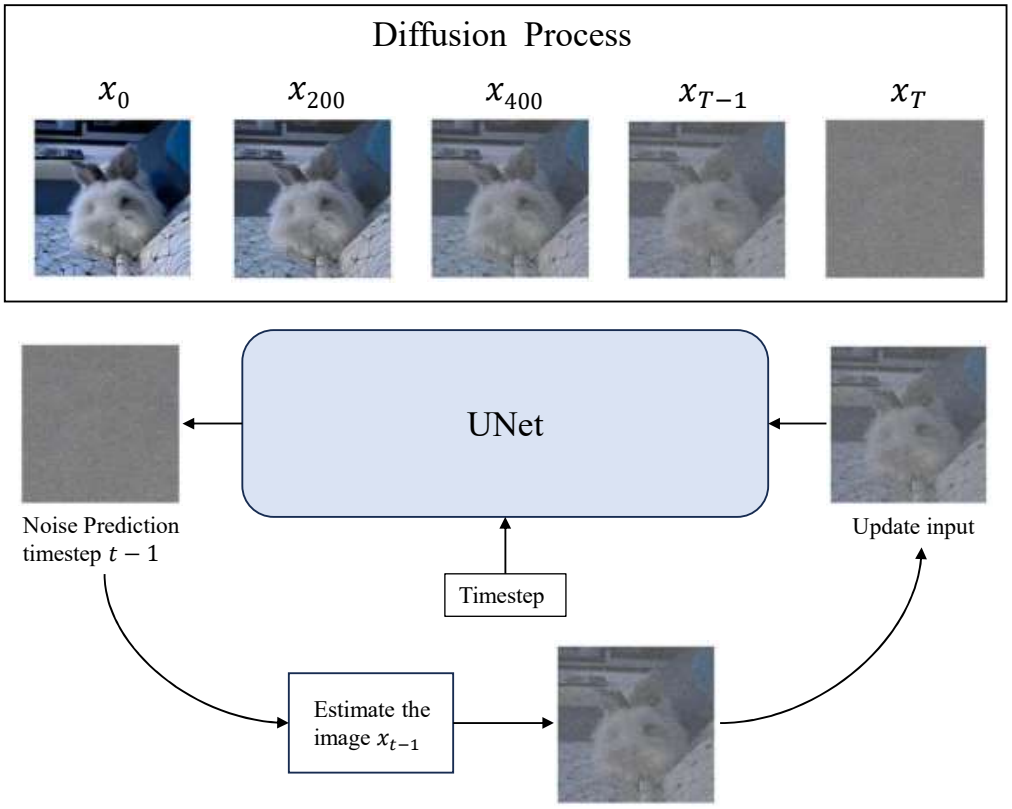


Figure 2.1: The diffusion process in diffusion models. Diffusion models recover data from Gaussian noise by progressively removing prediction noise at each time step using a series of Markov chains. x_t is the image in the timestep t, T is the total timestep.

the training process of the diffusion models is relatively complex, requiring multiple diffusion steps and sampling iterations, resulting in a long training time. Second, the generation speed is slow and images cannot be generated in real-time. In addition, parameter selection has a great influence on the generated results. Careful adjustment and selection of appropriate parameter settings are required.

2.2 Conditional Image Generation

Conditional Image generation is an image generation technique that controls the generation process by introducing conditional information to generate images that match given conditions. These conditions can be textual captions, semantic labels, auxiliary images, sketches, etc. Conditional image generation can be used in a variety of tasks, including image editing[45], image generation[43], image transformation[46], etc.

Compared with traditional unconditional generation methods, conditional image generation introduces additional input conditions, enabling the generator to generate images with specific properties based on conditional information. In previous studies, conditional image generation based on GANs is a common method, which combines the generative ability of GANs and conditional information to generate images with specific properties. A proposed method[47] addresses the challenge of conditional image-to-image translation, where the objective is to convert an image from the source domain to the target domain, taking into account a given image in the target domain as a condition. They address this issue using unpaired data, employing GANs and dual learning techniques. Lifelong GAN[48] delves into the lifelong learning challenge for generative models, allowing a trained network to adapt to new conditional generation tasks without erasing knowledge of previous tasks, all while relying solely on access to the training data for the current task. In Lifelong GAN, knowledge distillation is utilized to transfer accumulated knowledge from previous networks to the new network, allowing for image-conditioned generation tasks in a lifelong learning context.

Different from the mature conditional image generation of GANs, the conditional image generation of diffusion models is still under exploration. ControlNet[49] puts forward a neural network structure designed to control pre-trained diffusion models, facilitating the integration of supplementary input conditions. Two copies of a large diffusion model are duplicated by ControlNet with pre-trained weights. The capabilities learned from a vast number of images are preserved in the locked copy, whereas the trainable copy undergoes fine-tuning on task-specific datasets to acquire conditional control.

Another approach encodes conditional information into latent embeddings, which are then mapped to intermediate layers of U-Net via cross-attention layers. In this way, GLIGEN[50] implements bounding boxes, reference images, and keypoints as conditional information to control image generation based on the latent diffusion model. This new method endows new grounding controllability over existing text-to-image diffusion models.

The advantage of conditional image generation is that it provides greater control and customizability, enabling users to generate images with specific properties as desired. However, conditional image generation also faces some challenges, such as the accuracy and completeness of conditional information, the diversity, and scale of training data, etc.

2.3 Sketch-Based Image Generation

As an intuitive way, sketches are widely used in various studies. Generating lifelike images from freehand sketches poses challenges in the fields of computer graphics and computer vision.

Prior methods either require precise edge maps or depend on retrieving existing images. SketchyGAN[6] introduced a new GAN technique capable of generating lifelike images from 50 classes, encompassing airplanes, horses, and couches. This method did not retrieval the images at test time and directly copy input edges. Contextual GAN[51] use sketches as weak constraint, where sketches play a crucial role in providing the image context necessary for completing or generating the output image. This model facilitates straightforward and efficient learning of the joint distribution within the same image-sketch space, thereby avoiding the complexities involved in cross-domain learning. EdgeGAN[7] presented the first framework to generate realistic images from freehand scene sketches. This method also provided a large-scale freehand sketch dataset called SketchyCOCO to facilitate the research of object-level image generation from freehand sketches.

Recently, sketch-guided image generation based on diffusion models has also been increasingly developed. A compact per-pixel MLP network, called LGP[52], has been introduced to convert latent features of noisy images into spatial maps. With this method, the user gains intuitive control for the input sketches and semantic control for the output images. However, the quality of the results in this model may decreases on intricate scenes containing a blend of unclear and uncertain meanings. In DiffFaceSketch[14], a Multi-Auto-Encoder (AE) is employed to encode diverse sketches capturing different facial regions, converting them from pixel space to a latent space. This approach allows the model to compress the sketch input's dimensions while

retaining the geometry-related details of local facial features.

The advantage of the sketch-based image generation method is that it intuitively guides the image generation process. Sketches as input provide precise control over the desired image structure and features. However, there are still some issues, including the accuracy and diversity of the sketches, and how to maintain the details and authenticity of the generated images.

2.4 Two-Stage Image Generation

Two-stage models decompose a complex task into two independent stages to solve. These staged approaches are commonly employed to address intricate tasks, making the tasks easier to implement, and improving task efficiency and performance.

In the fields of computer graphics and computer vision, two-stage models have achieved remarkable results in many tasks. DualSlide[53] introduced a two-stage interface system based on sketches, encompassing global and local stages, to offer slides design by image retrieval and user guidance. During the global stage, a heat map canvas was presented to illustrate the distribution of all slide layouts within a dataset. In the local stage, comprehensive references and guidance can be given to aid in the design of slide content. Through a two-stage design strategy, DualMotion[54] facilitates the combination of global motions of lower limbs and local motions of upper limbs from a database. In the global design stage, users have the flexibility to initiate motion design by sketching a preliminary trajectory of body or lower limb movement. Subsequently, in the local design stage, users further refine the upper limb motions by sketching multiple relative motion trajectories.

In image generation, two-stage models are widely applied to conditional image generation for generating high-quality and diverse images. A two-stage drawing assistance system proposed by DualFace[13] provides users with global and local guidance. The global guidance assists users to draw contour lines for faces, while the local guidance aids in drawing facial part details. AniFaceDrawing[55] adopted a latent space exploration method of StyleGAN[22] with a two-stage training method to generate high-quality anime portraits. In the initial stage, an image encoder of StyleGAN is utilized and trained by AniFaceDrawing as a teacher encoder. In the subsequent stage, it emulated the drawing process without the need for any supplementary data. StackGAN[25] is a two-stage image generation model that can generate realistic images from text descriptions. The first stage produces a rough image layout, and the second stage refines and adds details to produce a realistic image.

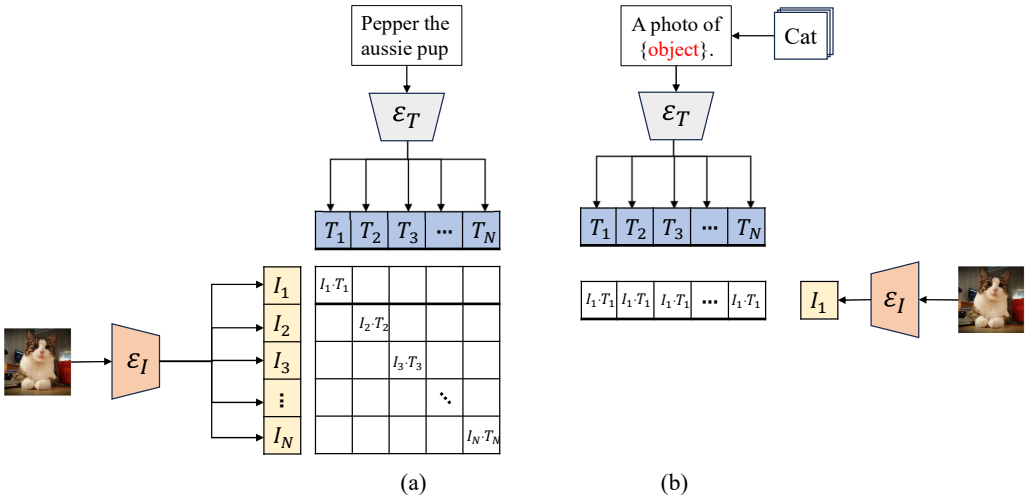


Figure 2.2: The framework of CLIP[2]. (a) CLIP co-trains an image encoder and a text encoder simultaneously to correctly associate a batch of (image, text) training examples and (b) the zero-shot prediction. Where $\mathcal{E}_I$ is the image encoder, $\mathcal{E}_T$ is the text encoder, I_n is the image tokens, and T_n is text tokens.

Two-stage models have achieved significant success in various computer vision and computer graphics tasks. In this thesis, we aim to apply the two-stage method to address the issue of position control in the image generation process of the diffusion model. We decompose the generation process into two stages. Position guidance is provided to the model from sketches in the first stage, and a pre-trained text-to-image diffusion model is utilized to generate images in the second stage.

2.5 Contrastive Language-Image Pre-training

Contrastive Language-Image Pre-training (CLIP) is a powerful vision-language pre-training model proposed by OpenAI[2]. On a dataset of large scale image-text pairs gathered from the internet, CLIP achieves state-of-the-art visual concepts. Following pre-training, the model leverages natural language to refer to learned image representations, facilitating zero-shot migrated to following tasks. The primary goal of CLIP is to achieve joint representation learning of images and texts within a unified framework, enabling the model to comprehend the relationship between vision and language.

As shown in Figure 2.2, the main structure of CLIP consists of a text encoder and an image encoder, which compute embeddings for text and images

respectively. The text encoder employs the Transformer[34], while the image encoder uses two models, ResNet[56] and Vision Transformer (ViT)[57]. The encoded image and text vectors are mapped to a joint multimodal space, producing new vectors that can be directly compared (image vector I_e and text vector T_e, e is a natural number). Subsequently, the similarity between the text vector and the image vector is then calculated to predict whether they form a pair.

The unique feature of the CLIP lies in its ability to automatically learn the correlation between vision and language without the need for image-text pairing or alignment. It achieves this through unsupervised learning, enabling CLIP to handle data from different languages and domains, making it highly versatile for multilingual and multimodal tasks.

CLIP has received extensive recognition and demonstrated impressive performance in diverse tasks related to computer vision (CV) and natural language processing (NLP). It excels in tasks such as image retrieval[58] and image generation[1].

2.6 DeepLab-V2

DeepLab-V2 was proposed by Chen et al initially. for the task of semantic image segmentation using deep learning[59]. Sketchyscene customized DeepLab-v2 for segmenting scene sketches[60]. DeepLab-V2 won first awards in the PASCAL VOC 2012 Semantic Segmentation Challenge, demonstrating its excellent performance.

DeepLab-V2 is a Fully Convolutional Network (FCN) based model that transforms a classification model into a segmentation model by substituting the final fully connected layers with fully convolutional layers. DeepLab-v2 encompasses three crucial components: atrous spatial pyramid pooling (ASPP), atrous convolution, and the application of fully-connected Conditional Random Field (CRF) for post-processing. Traditional convolutional kernels have closely arranged sampling within their receptive fields, while the atrous convolution technique introduces holes (or dilations) within the kernels, making the sampling sparser and thus enlarging the receptive fields. This allows the model to obtain more extensive contextual information and comprehend the semantic content of the image. The main idea of ASPP is to perform multi-scale information extraction using different dilation rates. It is analogous to feature extraction at different receptive fields, and the owned features are then fused to obtain a more comprehensive contextual understanding. In the post-processing stage of segmentation results, DeepLab-V2 utilizes a CRF layer to smooth the predicted results. The CRF layer in-

troduces conditional dependencies between class labels, reducing small segmentation errors and making the segmentation results more coherent and refined.

Chapter 3

Conditional Generation with Latent Diffusion Model

In this section, we first introduce the preliminaries of the diffusion model in Section 3.1. We then introduce the method of the latent diffusion model (LDM) in Section 3.2 and the attention mechanism in Section 3.3. In Section 3.4, we discuss the effect of the attention maps in the latent diffusion model. In Section 3.5, we introduce the backbone of LDM. Our framework and implementation details will be discussed in Chapter 4.

3.1 Diffusion Model

Compared with GANs, the diffusion model excels in the field of image generation, achieving superior results. For example, Dhariwal et al. showed that diffusion models get FID values on ImageNet 128×128 (2.97 value), ImageNet 256×256 (4.59 value), and on ImageNet 512×512 (7.72 value). The results match BigGAN-deep, the latest high-resolution, high-quality image generation GAN[10], even with a minimal number of 25 forward process per sample, and diffusion models preserve a more comprehensive coverage of the distribution[61]. However, the diffusion model requires repeated iterative calculations with increased training and reasoning costs. They also reported that estimated takes $150 - 1000$ days to train on NVIDIA Tesla V100. Thus, we aim to utilize the pre-trained diffusion model to generate position-controlled scene images without fine-tuning or retraining.

The working principle of the diffusion model is to learn the information decay caused by noise and then use the learned patterns to generate images[29]. As shown in Figure 3.1, diffusion models contain both forward diffusion process and reverse denoising (inference) processes. The forward

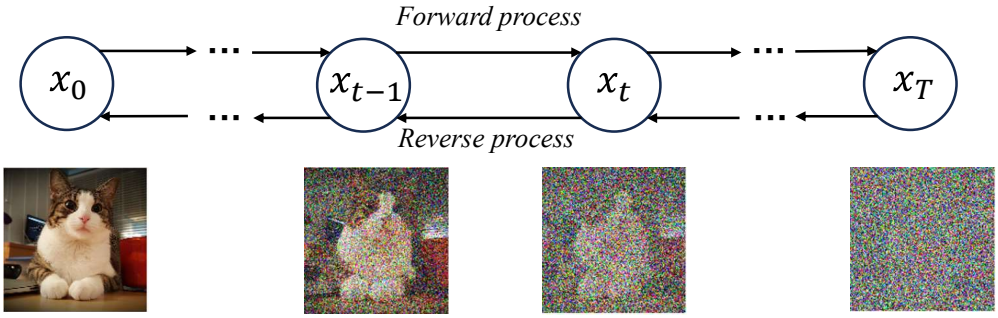


Figure 3.1: In the forward process, noise is introduced into the data samples gradually. In the reverse process, the noise sample is gradually denoised to generate images.

process follows the concept of a Markov chain and turns the input image into Gaussian noise. Given a data sample x_0, the Gaussian noise is progressively Increased to the data sample during T steps in the forward process, producing the noisy samples x_t, where the timestep $t = \{1, \ldots, T\}$. As t increases, the distinguishable features of x_0 gradually diminish. Eventually when $T \to \infty$, x_T is equivalent to a Gaussian distribution with isotropic covariance. In addition, the inference process can be understood as a sequence of denoising autoencoders with same weights $\epsilon_\theta(x_t, t)$ (ϵ_θ is typically implemented as U-Net[5]), which are trained to forecast denoised images of their corresponding inputs x_t. The corresponding objective function can be written as follows:

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{3.1}$$

where t is uniformly sampled from $\{1, \ldots, T\}$, ϵ is the sample noise from normal distribution, $\mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}$ means the evidence lower bound (ELBO) associated with the diffusion model[1].

3.2 Latent Diffusion Model

The Latent Diffusion Model (LDM) [1] proposed a method of performing the diffusion process on the latent space, which can greatly reduce the computational complexity with high-quality image results. As shown in Figure 3.2, LDM can generate detailed images, and it also performs well on high-resolution image generation tasks (such as landscape image generation and megapixel images). In addition, LDM also can conduct unconditional image generation, inpainting, and super-resolution tasks.

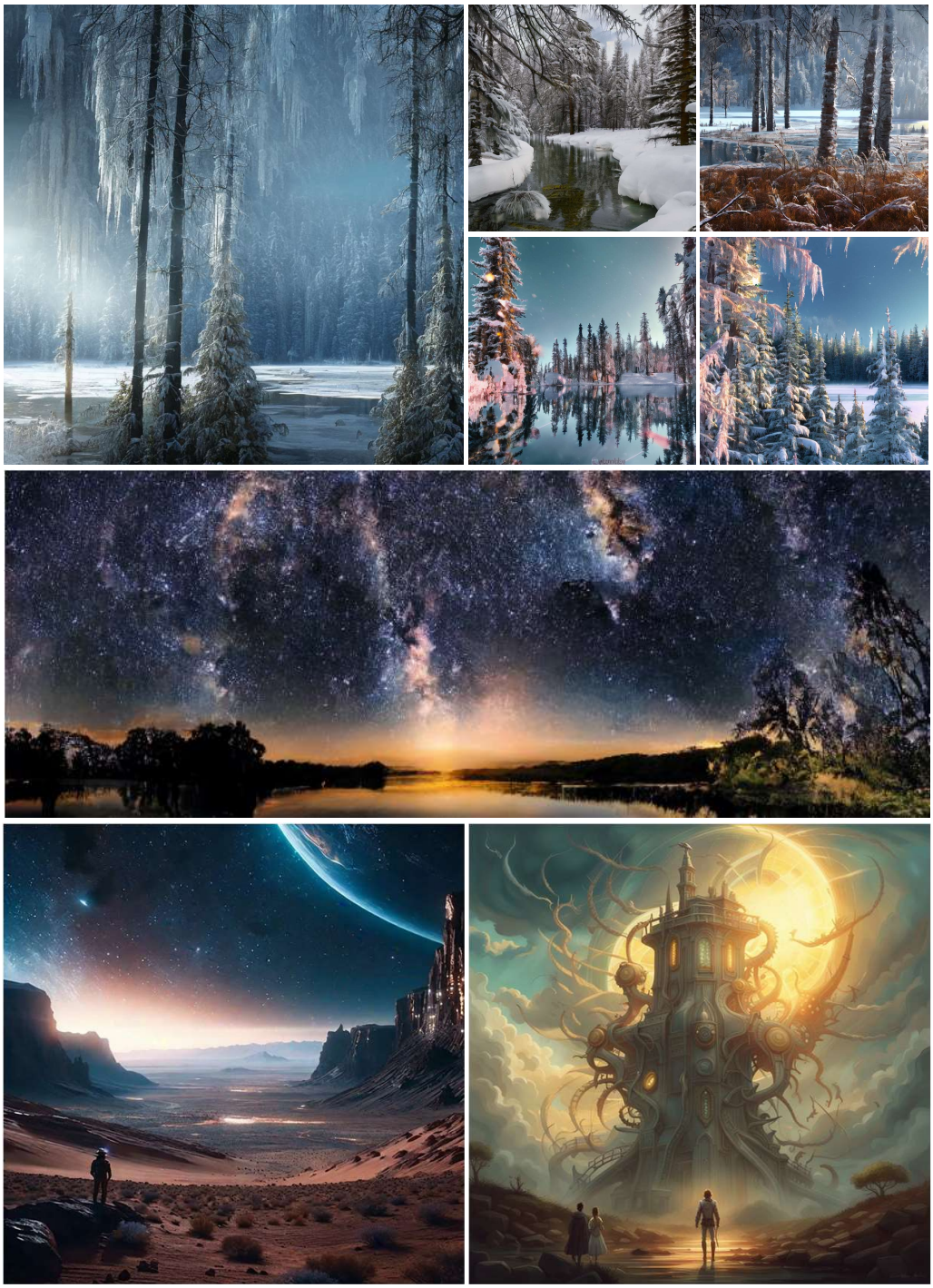


Figure 3.2: Landscape images generated by LDM (version: XL). The images in the first row were generated with 768×768 resolution. The generated images also can generalize to larger resolutions (in the second row: 1024×384; in the third row: 1024×1024 (left), 2048×2048 (right)).

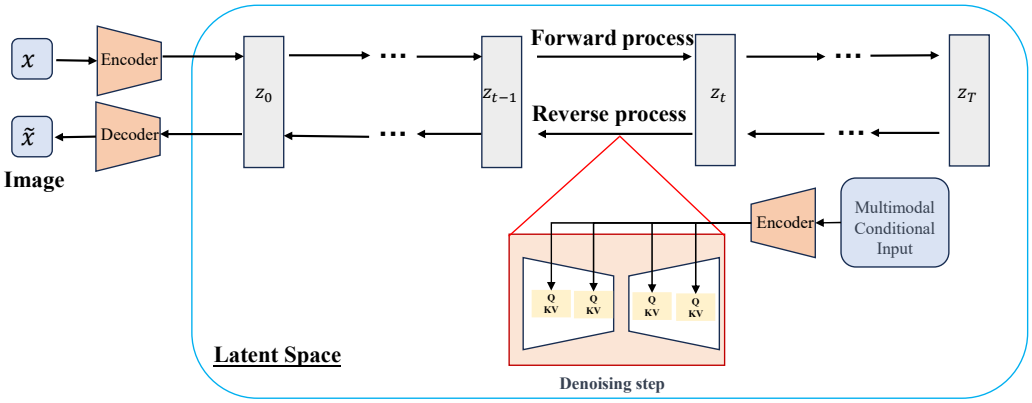


Figure 3.3: The framework of the latent diffusion model, which is proposed by Rombach et al[1]

.

The difference between LDM and the traditional diffusion model[29] is that LDM does not directly operate on the images but operate in the latent space. LDM calls this method perceptual compression. The perceptual compression model is based on previous work[62] and composed of an autoencoder that underwent training through a patch-based adversarial objective and a blend of a perceptual loss. LDM reduces the dimensionality of the data by projecting it into a low-dimensional, efficient latent space, in that high-frequency, imperceptible details are abstracted away. Perceptual compression is typically employed to reduce computational complexity, save storage space, and improve the efficiency of model training and inference.

The framework of LDM is illustrated in Figure 3.3[1]. LDM trained an AutoEncoder, including an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. After the image x is compressed by the encoder $\mathcal{E}$ to latent representation z, the diffusion process is performed on the latent representation space. LDM has a similar diffusion process to the standard DM. Finally, LDM infers the data sample z from the noise z_T and $\mathcal{D}$ restores the data z to the original pixel space and gets the result images $\widetilde{x}$.

Specifically, given an image $x \in \mathbb{R}^{H\times W\times 3}$ with height H, wigth W in RGB space , LDM first utilizes an encoder $\mathcal{E}$ to encode the image x into a latent representation space:

$$z = \mathcal{E}(x) \tag{3.2}$$

where $z \in \mathbb{R}^{h\times w\times c}$ with height h and width w, the constant c represents the number of channels. The encoder $\mathcal{E}$ downsamples the image by a factor $f = H/h = W/w$ (LDM discussed the impact of different c and f on the model

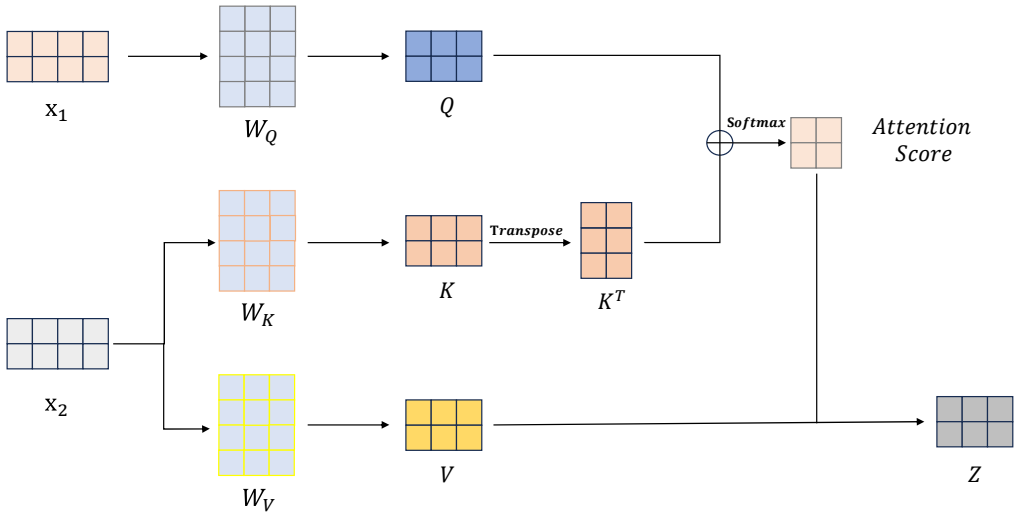


Figure 3.4: The framework of cross-attention mechanism. Where X_1 and X_2 are different inputs (such as text and sketch), Z is the output, W_Q, W_K, and W_V are learnable projection matrice in LDM. The cross-attention mechanism calculates the attention score according to K and Q, and applies V to the attention score to obtain the final output.

in the appendix). Then $\mathcal{D}$ recover the image from the latent representation space:

$$\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x)) \tag{3.3}$$

Since a pre-trained perceptual compression model is introduced in LDM, which includes $\mathcal{E}$ and $\mathcal{D}$, the model can obtain the noise sample z_t in latent space by encoder $\mathcal{E}$ during training (corresponds to the x_t in pixel space) and conduct the forward diffusion process in the latent representation space. Thus, the Equation 3.1 can be written as follows:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t\right)\|_2^2\right] \tag{3.4}$$

3.3 Attention Mechanism

The attention mechanism is a key component in computer vision tasks, such as image recognition and object detection. It allows models (such as Transformer[34], BERT[63] and GPT[64]) to focus on specific regions or features of an input image that are considered important for the task. Self-attention and cross-attention are two variations of the attention mechanism commonly used in deep learning models. Especially, self-attention is used to

calculate the relationship between elements in the input mode, and cross-attention is designed to calculate the relationship between elements in multimodal input. The main difference between them is the source of queries (Q), keys (K), and values (V) used to calculate the attention score.

We focus on cross-attention because cross-attention allows multimodal inputs (such as text, sound, and image). As shown in Figure 3.4, the cross-attention asymmetrically combines two embedding modes of the same dimension, while one mode is used for calculating Q, and the other mode is used for calculating K and V.

LDM can be used to explore conditional image generation, which is mainly obtained by expanding the conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$. y is the conditional information that controls the process of image generation.

Specifically, LDM implements $\epsilon_\theta(z_t, t, y)$ by adding a cross-attention mechanism to the U-Net backbone network. In order to easily introduce various types of conditioning y (such as text, layout, sketch, etc.), LDM introduces a domain-specific encoder τ_θ, which is used to map y to an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, where M is index dimensionality (such as $M = 50176$ for 224×224 ImageNet images[65]) and d_τ is channel dimension of τ_θ.

Finally, LDM integrates the conditional information into the middle layer of U-Net through cross-attention layers mapping. The implementation of the cross-attention layer is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{3.5}$$

, with $Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y)$ of dimension d . where $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ is an intermediate representation of U-Net, N is the latent's index dimension. W_Q, W_K, and W_V are learnable projection matrice in LDM.

In this case, the Equation3.4 can be deduced as:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{3.6}$$

3.4 Attention Maps

LDM proposed the cross-attention method to achieve multi-modal training to realize the conditional image generation task. The CLIP encoder encodes provided text as an embedding sequence during a generation process, which is subsequently processed into keys and values K, V. LDM uses latent seeds to control the diversity of generated samples. The generator creates the initial

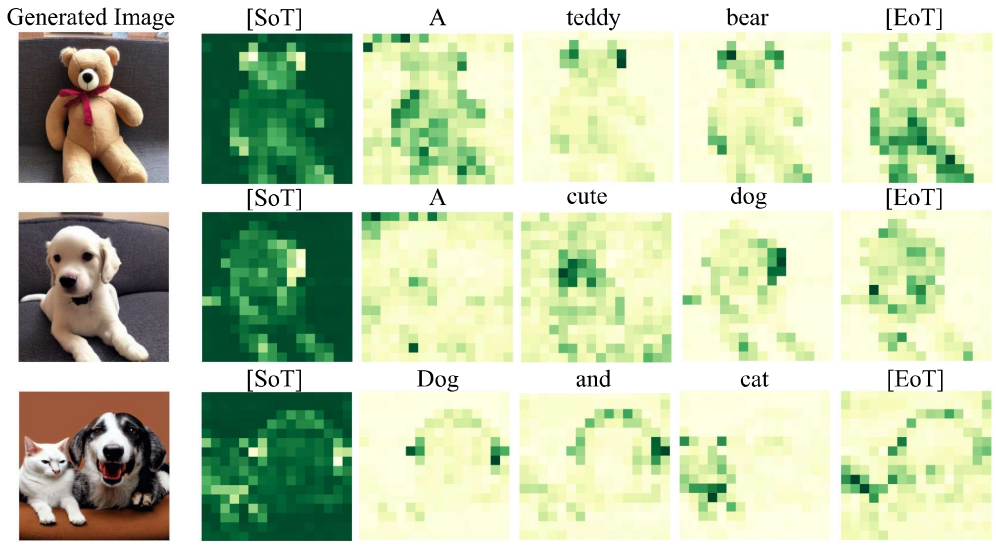


Figure 3.5: The text tokens compose of a start token [SOT], text content, and many padding tokens [EOT], and the attention maps contain the object locations corresponding to the text tokens[3].

latent noise according to the latent seed, and the noise is encoded as visual tokens and calculated for Q. The attention maps M can be calculated as follows,

$$\mathrm{M} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{3.7}$$

The attention map M controls the spatial distribution of values V, which contains rich semantic information.

Chen et al.[3] explored the cross-attention maps and gave insights. As shown in Figure 3.5, all of the text tokens compose of a start token ([SOT]) and many padding tokens ([EOT]). The padding tokens ensure that the input prompts with different lengths are mapped into the tensor to get the text tokens with the same length. Upon analyzing the [SOT] and [EOT] in detail, it was discovered that the cross-attention maps of these tokens also possess meaningful semantic and spatial information. The key discovery is that cross-attention maps have a predominant influence on determining the objects' position of the generated images.

Hertz et al.[4] also reached a similar conclusion - The spatial arrangement and shapes of objects in the generated image are contingent on the cross-attention maps. As shown in Figure 3.6, the pixels exhibit a stronger affinity towards the words that describe them. For example, the pixels of the bear ex-

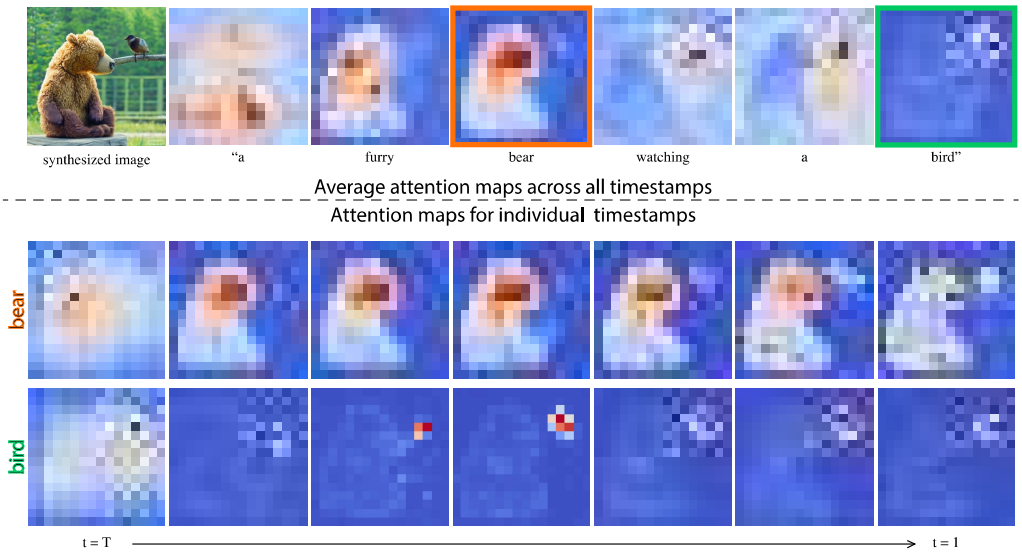


Figure 3.6: The attention maps could place a higher focus on the homologous objects, and the objects' positions have been determined by the attention maps early in the diffusion process[4].

hibit correlation with the word “bear”. In addition, the cross-attention maps of adjectives affect the image representation ability of corresponding nouns. Interestingly, The image's structure is established during the initial stages of the diffusion process. Most importantly, the degree to which attention is injected into the diffusion process affects the quality of the generated results. The greater the number of diffusion steps in which cross-attention injection is applied, the stronger the consistency between images and the conditional information. However, applying the injection throughout all diffusion steps does not necessarily achieve the optimal result.

3.5 U-Net

U-Net is a deep learning model used for semantic segmentation tasks[5]. Its key feature is the adoption of an encoder-decoder architecture, which efficiently performs semantic segmentation on images while requiring fewer training samples.

As shown in Figure 3.7, the encoder of U-Net consists of a series of convolutional layers and pooling layers, progressively reducing the input image size and extracting high-level feature representations. The decoder comprises a series of transposed convolutional layers, aimed at gradually restoring the

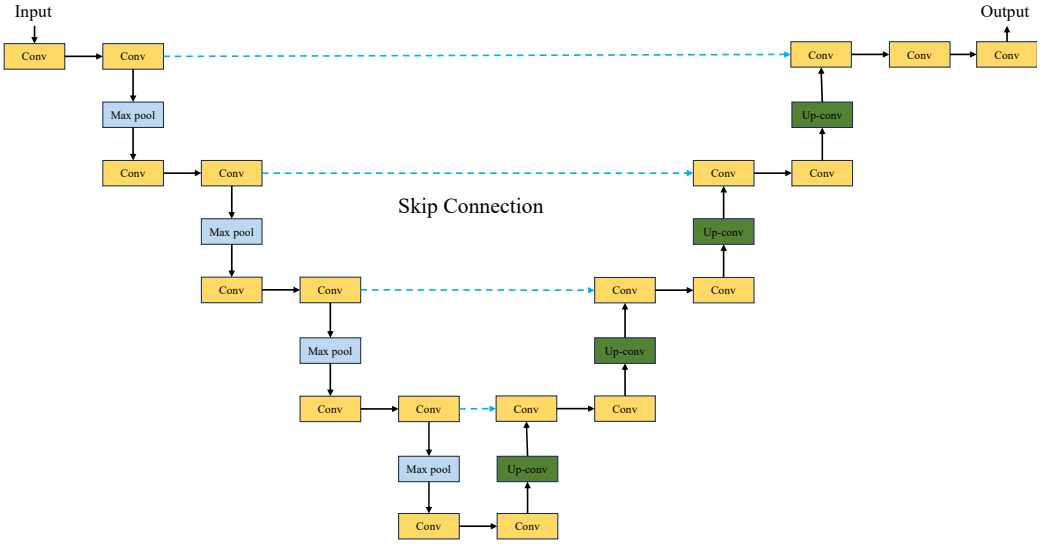


Figure 3.7: The architecture of U-Net (an illustration for the lowest resolution of 32x32 pixels.)[5].

original image size and generating segmentation results. The decoder utilizes transposed convolutions for upsampling and incorporates skip connections, connecting features from the encoder to the decoder. This skip connection design enables the transfer of information between the encoder and decoder, aiding in preserving more contextual information and improving segmentation accuracy.

U-Net in LDM adds a time embedding module and spatial transformer (cross-attention) modules to the basic encoder-decoder U-net. As shown in Figure 3.8 Time embedding is the process of mapping time information to a continuous vector space, allowing the model to learn and utilize temporal relationships. LDM requires multiple iterations to iteratively predict noise, using time embedding to encode time information into the network, enabling U-Net to predict more appropriate noise at each iteration. LDM utilizes a cross-attention module to control the fusion and interaction between textual and image information. Specifically, the cross-attention module guides U-Net to align certain regions of the noise matrix with specific information from the text.

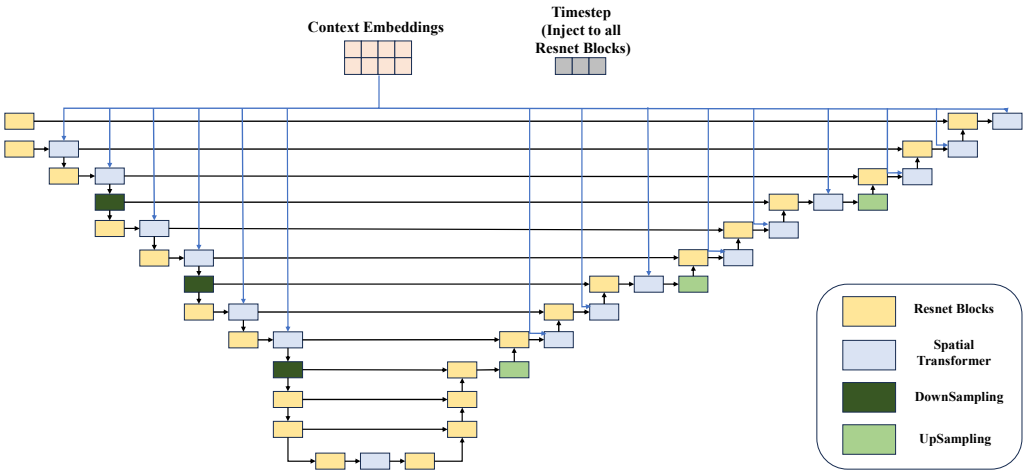


Figure 3.8: The illustration of time-conditional U-Net with attention mechanism.

Chapter 4

Sketch-Guided Image Generation

We discuss the detailed composition of our proposed two-stage scene image generation model with position control in this chapter. We first give an overview of our proposed model in Section 4.1. We then introduce the SketchyCOCO dataset in Section 4.2. We also introduce the two stages in our proposed model. In the feature extraction stage, we utilize instance segmentation to extract object locations and labels from sketches and encode them as position guidance of the generation process (introduced in Section 4.3). In the image generation stage, the pre-trained LDM generates images according to the input text prompts, where the objects' positions will follow the position guidance of the sketches (introduced in Section 4.4). We continue to deduce the cross-attention formula from Section 3.3 and introduce the variations of the formula in our method.

4.1 Framework Overview

Different types of conditions have been considered in the latent diffusion model, such as the example image, layout, and key points. Sketches can contain information about the layout and composition of an image. It can indicate the relative position and scale of major elements, as well as the relationship between them. Layout and composition determine how the individual elements in an image are arranged, affecting the overall look of the resulting image. Additionally, sketches can provide information about objects' scale and proportional relationships in an image. It can indicate objects' scale, aspect ratios, and relative proportions between them. This is important to ensure that the resulting images are properly scaled and realistic. In

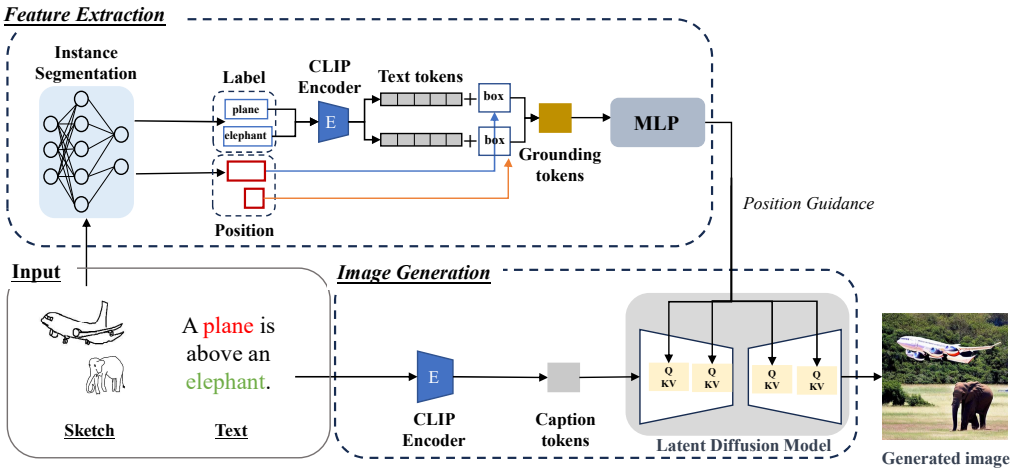


Figure 4.1: The framework of our model. The model first extracted the sketch's features and introduced them into the attention layers with caption tokens to generate the images.

cases where detailed generation is difficult to easily control with text prompts alone, we would like to directly intervene in the attention mechanism with user-specified input sketches to generate corresponding position-controlled images.

Our goal is to generate high-quality scene images with the position guidance of human-drawn scene sketches. Our method set text prompt and sketch as inputs, where the text prompt provides the global description and the sketch provides the conditional control. Our proposed two-stage framework uses a text-to-image diffusion model that has been pre-trained to generate controllable images without additional training or fine-tuning. In this two-stage model, the feature extraction stage constrains a set of constraints (position, label, etc.) extracted from the sketch and feeds them into attention layers to influence the position and shape generation in the early stage of the diffusion process. The image generation stage leverages the generative capabilities of the latent diffusion model to generate images following the position guidance from the feature extraction stage.

As shown in Figure 4.1, we use both a sketch and a text prompt as inputs. The text prompt is encoded into text embeddings by the encoder of CLIP, called caption tokens in LDM, providing global semantic information for LDM. The sketch serves as a conditional input and undergoes instance segmentation. We employ the pre-trained DeepLab-V2 model to obtain corresponding labels and bounding boxes. The labels are encoded into temporary text tokens by the encoder of CLIP and combined with the upper left

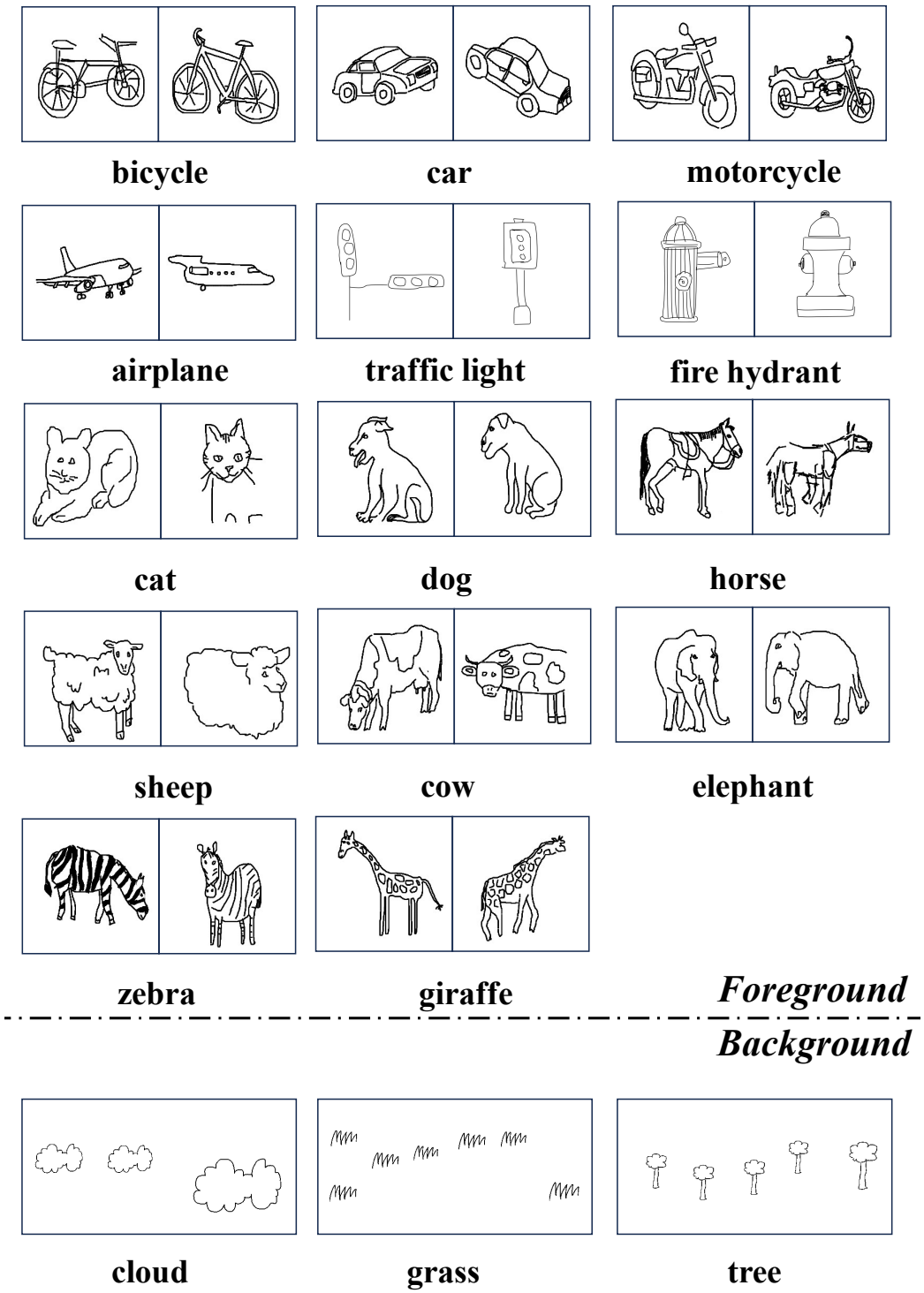


Figure 4.2: The sketches in SketchyCOCO dataset, which include 14 categories of objects and 3 categories of background freehand sketches.

and lower right coordinates of the bounding boxes to form the final grounding tokens. Finally, the grounding tokens are passed through an MLP network and inputted into the attention layers of the LDM to provide the position guidance for image generation.

4.2 SketchyCOCO Dataset

Many sketch datasets have been proposed, providing a crucial foundation for the research and development of sketch-related algorithms. The QuickDraw[66] dataset is a comprises of 50 million vector sketch images of 345 classes. It contains a large number of freehand sketches covering a wide range of object categories and concepts. However, since the sketches are drawn by users within 20 seconds, some of the sketches in the dataset have lower quality issues, such as being incomplete, blurry, or distorted in shape. Additionally, the QuickDraw dataset does not include scene sketches.

SketchyScene[60] presented the first large-scale dataset of scene sketches, with encompasses of over 29,000 scene sketches, (all objects in the scene sketches come with grounding semantic and instance masks.) more than 7,000 pairs of scene templates and their corresponding photos, and over 11,000 object-level sketches. The sketches in SketchyScene are generated or synthesized by users based on reference images, and all the sketches have fine details and high quality. However, the majority of individuals are not professionally trained artists, making it challenging for them to draw intricate scene sketches, particularly when objects are in various shapes and poses.

In this case, we focus on the SketchyCOCO[7] dataset, a large-scale dataset of hand-drawn scene sketches based on MS COCO Stuff dataset[67], proposed to study sketch-based image understanding, retrieval, and generation tasks.

As shown in Figure 4.2, SketchyCOCO collects 20198 triplets of foreground examples <sketches, images, edge maps> in 14 categories (airplane, cat, giraffe, zebra, dog, elephant, fire hydrant, horse, bicycle, car, traffic light, cow, motorcycle, sheep), 27683 pairs of background examples <sketches, images> in 3 categories (cloud, grass, tree), 14081 pairs of intermediate products <foreground image or background sketch, scene image>, 14081 pairs of scene sketch-image examples, and the grounding segmentation for 14081 paired scene sketches. The introduction of the SketchyCOCO dataset provides a significant resource and benchmark for sketch-based research. It enables researchers to explore the connection between freehand sketches and images. The dataset's diversity and richness make it a crucial research asset in computer vision.

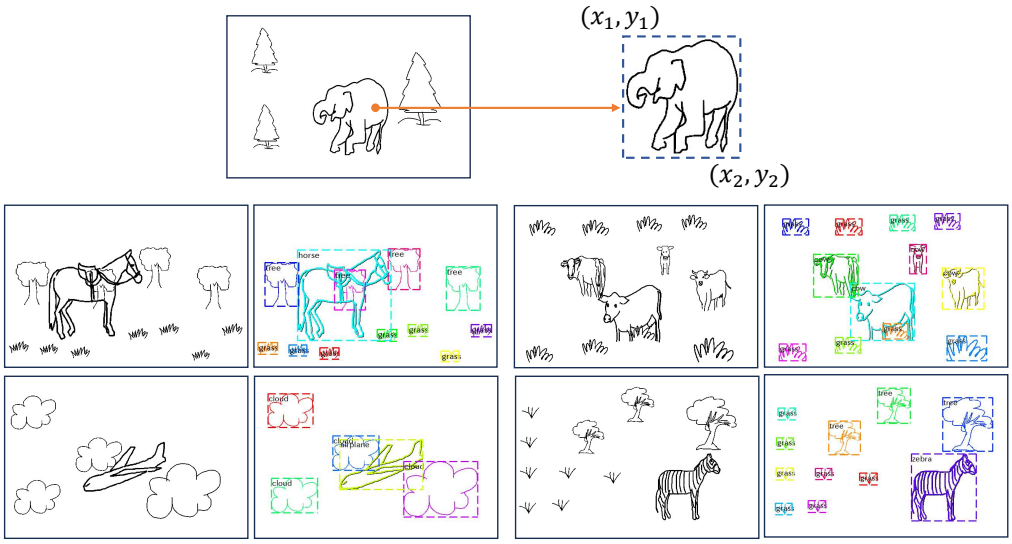


Figure 4.3: The visualized results of feature extraction stage. We divide the sketches into labels and bounding boxes and capture the coordinates of the top-left and bottom-right corners of the bounding boxes.

4.3 Feature Extraction Stage

In the feature extraction stage, we focus on extracting position information from the sketch for position control in the conditional generation. Inspired by SketchyScene[60], We employ the segmentation model based on DeepLab-v2 as the segmenter $\mathcal{S}$ to complete the instance segmentation, which is customized for segmenting scene sketches. Therefore, for the input sketch x_s, it can be expressed as $\mathcal{S}(x_s)$.

As shown in Figure 4.3, after segmentation, the corresponding bounding boxes and labels of the objects can be obtained. The labels l represent the corresponding names, such as "cow", "tree", and "airplane". The bounding boxes b represent the coordinates$[x_1, y_1, x_2, y_2]$, where (x_1, y_1) represents the top-left coordinate and (x_2, y_2) represents the bottom-right coordinate. The segmentation can be expressed as

$$(l, b) = \mathcal{S}(x_s) \tag{4.1}$$

The labels will be encoded by the CLIP text encoder as the text tokens. The text tokens will be combined with coordinates as grounding tokens and inputted to LDM for conditional control. Thus, We define our model as a composition of the caption and grounding sketch:

$$I = (c, e) \tag{4.2}$$

$$e = \mathcal{S}(x_s) \tag{4.3}$$

where I is the generated image, c is the text caption and e is the grounded tokens. e also can be expressed as $e = (l, b)$.

4.4 Image Generation Stage

In the image generation stage, the pre-trained LDM generates images according to the input text prompts with position guidance from the feature extraction stage. As mentioned above in Section 3.3, the initial stages of the diffusion process already establish the shape and position of the image. Therefore, during the image generation stage, we use position guidance to influence the position generation of objects on the attention maps in the initial stages of the forward diffusion process. Subsequently, we employ LDM to generate images only based on text prompts.

4.4.1 Cross-Attention Layer

We first introduce the cross-attention and deduce the cross-attention formula in our model. Cross-attention is a commonly used attention mechanism for the processing of multimodal inputs. The application of cross-attention can help the model to build correlations between different inputs, so as to understand the semantic relationship between them.

During the diffusion process, the cross-attention mechanism is employed to direct the image generation process, enabling the model to understand and process images at different scales and levels. The cross-attention mechanism combines the concepts of attention mechanism and cross-layer connections. Its role is to establish efficient information transfer and interaction between different layers of the model.

As shown in Equation 3.7, LDM integrates the conditional information into the middle layer of U-Net through cross-attention layers mapping. In the original latent diffusion model, Q comes from visual tokens generated from latent seeds, and both K and V come from caption tokens in the text.

In our model, we kept K and V unchanged and still included the feature information in the caption. Inspired by the GLIGEN[50], We fuse the obtained grounding tokens and visual tokens as Q to query in the attention layers. Thus the Q can be expressed as

$$Q = v + \beta \times tanh(\gamma) \times e \tag{4.4}$$

where v is the visual tokens from the latent seeds, β is a gated parameter that will be introduced in Section 4.4.2 and γ is a learnable scalar.

4.4.2 Gated Parameter β

For a diffusion process with T time steps, we can set a fixed time step αT to divide the diffusion process, where α is a constant. When time step $t \leq \alpha T$, this indicates that the diffusion process is at an early time, at which point we condition the control via set $\beta = 1$. At this time, the Q of attention layers will be composed of grounding tokens e and original visual tokens v:

$$Q = v + tanh(\gamma) \times e \tag{4.5}$$

Therefore, we control the generation of positions early in the diffusion process.

When $t \geq \alpha T$, the model set $\beta = 0$. At this situation, we use the original generation ability of LDM for image generation. Thus, the Q of attention layers will be the original visual tokens v:

$$Q = v \tag{4.6}$$

Note that the model in this situation has nothing to do with the additional input conditions, and the model maintains the original generation ability.

In summary, since the degree to which attention is injected into the diffusion process affects the quality of the generated results, we divided the image generation stage into two steps by β:

$$\begin{cases} \beta = 1, & t \leq \alpha T \quad \text{Position guidance} \\ \beta = 0, & t > \alpha T \quad \text{Standard inference} \end{cases} \tag{4.7}$$

We think α should be a variable parameter. Since different values of α will affect the ability to generate, we explored the impact of different α on generation in the section 5.2. Users can choose the appropriate value of α according to their design intentions.

Chapter 5

Experiment and Results

We conduct qualitative and quantitative experiments to verify the image quality and sketch input consistency of our model's generated scene images. In Section 5.1 we introduce the implementation details of our experiment. We present the results of our qualitative evaluations (Section 5.2) and quantitative experiments (Section 5.3).

5.1 Implementation Details

Both stages of our model are implemented on the Ubuntu system, i7-13700KF CPU, and a single NVIDIA RTX4090 GPU. In the conducted experiments, we use the pre-trained LDM-V1.4[1] (that trained on the LAION-5B dataset[68]) as the proposed image generator. The LAION-5B dataset is a large-scale graphic and text dataset that consists of 5.85 billion CLIP-filtered pairs of image and text, including 2.3 billion image-English text pairs, 2.2 billion images, and the remaining 1 billion pairs are not limited to specific languages, such as image names. The hyperparameters are shown in Table 5.1. All of our sketch scene images from the SketchyCOCO dataset[7], which include 14 categories of objects and 3 categories of background freehand sketches.

In quantitative comparison with LDM, we use the 300 randomly sampled sketches in the SketchyCOCO dataset to generate 300 images for evaluation. We provided the corresponding text prompts manually. In quantitative comparison with GANs (pix2pix[8], SketchyGAN[6], and SketchyCOCO[7]), we utilize the 200 randomly sampled sketches with a single object in the SketchyCOCO dataset for evaluation. The average generation time cost for the images is 25.3 seconds.

Hyperparamter	
z-shape	32×32 × 3
T	1000
f	8
Noise Schedule	linear
Channels	320
Depth	2
Conditioning	CA
α	0.4
γ	0.6

Table 5.1: The hyperparameters used in the proposed model. z-shape is the dimension of latent space, diffusion steps T and factor f are introduced in Section 3.2, α and β are introduced in Section 4.4.

5.2 Qualitative Evaluation

To illustrate the difference between our model and the state-of-the-art mainstream text-to-image model, we use Latent Diffusion Model[1] as a reference to illustrate.

As shown in Figure 5.1, the state-of-the-art mainstream text-to-image method obtains global information and generates an image from the text, but it cannot further control the position information of the generated image. We also compare with other text-to-image models, as shown in Figure 1.1. Our model uses the scene sketch as additional supplementary information to control the position generation of the scene image. It is verified that our image has achieved a relatively good effect in terms of position control, and all desired objects can appear in the corresponding position.

Different values of α will affect how much conditional information is injected into the attention layers, and then affect the position information and generation ability of image generation. We tried to explore the influences by different α value settings in our model. As shown in Figure 5.2, When no additional conditional information is injected, the proposed model only generates images according to the text prompt information and according to the original generation ability. The position is only controlled in the early stage of the diffusion process, and the image can also be generated according to the bounding boxes and labels of the sketches. Objects in the generated images can already appear in the correct position. In the third row, we consider completely anomalous images. The original pre-trained model misunderstands semantic information and loses objects. After injecting the

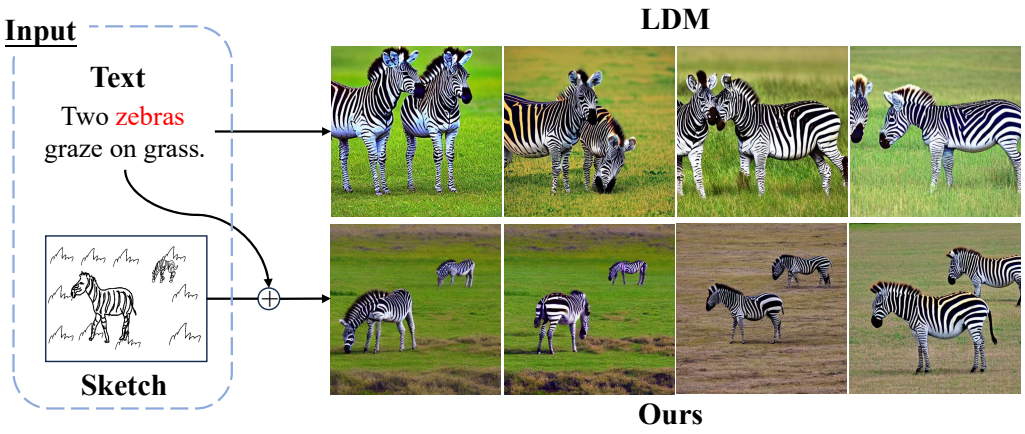


Figure 5.1: The difference of generated results between our model and LDM[1]. Our proposed method achieves the position guidance of an image generated by a pre-trained text-to-image diffusion model, such as Stable Diffusion [31].

position information of the sketch, the images are still generated according to the corresponding position, even though the proportion relationship between the "giraffe" and the "airplane" is completely abnormal.

As mentioned above, after noticing the phenomenon of object loss in the original LDM model, we conducted further exploratory experiments to verify that our model can help to improve the object loss issue. As shown in Figure 5.3, When there are multiple objects in the semantics, the pre-trained text-to-image LDM model will have situations of semantic loss and disordered positions. After adding our sketch as an auxiliary, the generated image can contain the correct number of objects and have the corresponding position information.

Due to the inclusion of two different inputs in our model (text and sketches), the text provides global semantic descriptions while sketches offer detailed control over specific objects. To explore the relative influence of these two inputs on image generation. As illustrated in Figure 5.4, under the given text prompt, we deliberately alter the positions of objects in the sketches to contradict the positional information described in the text. For example, we give the location description in the text prompt as "left" in the first row, but in the second, third, and fourth columns, we change the relative position in the sketches to "right", "top", and "bottom" respectively. The generated images consistently depict objects positioned according to our sketches, even if it contradicts the text.

We also conduct an experiment to verify the scale control of the condi-

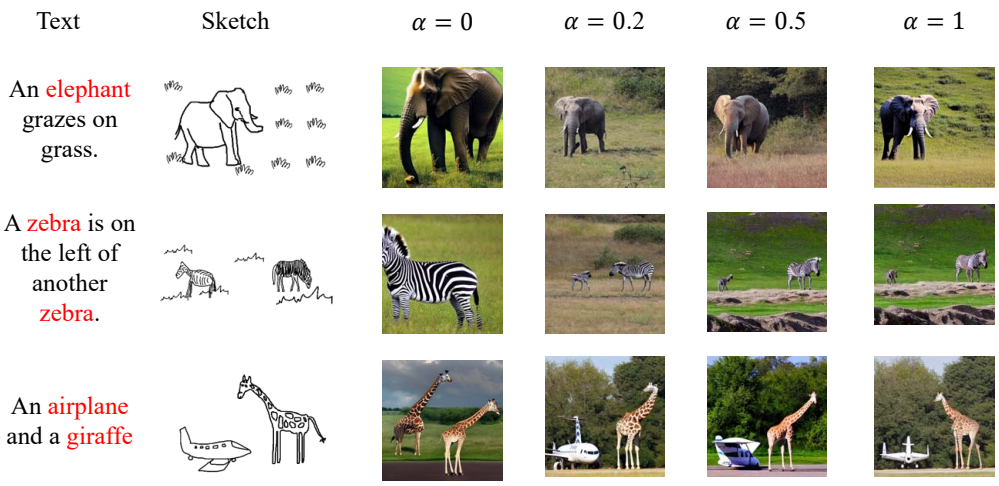


Figure 5.2: The generated images with different α values. In the first and second rows, we consider the condition with a single object and two objects. In the third row, we verified situations that are unreasonable in reality.

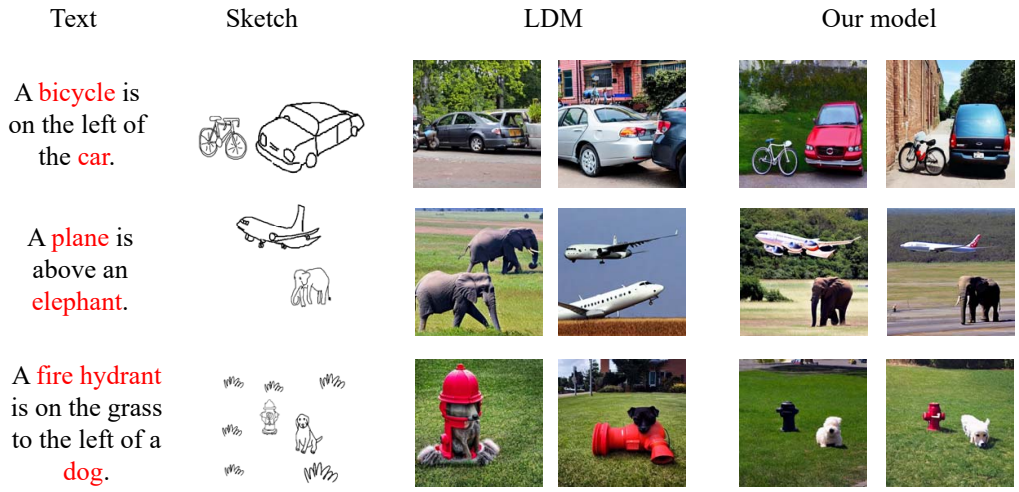


Figure 5.3: Our model can effectively improve the object loss issue that occurs in the original LDM model.

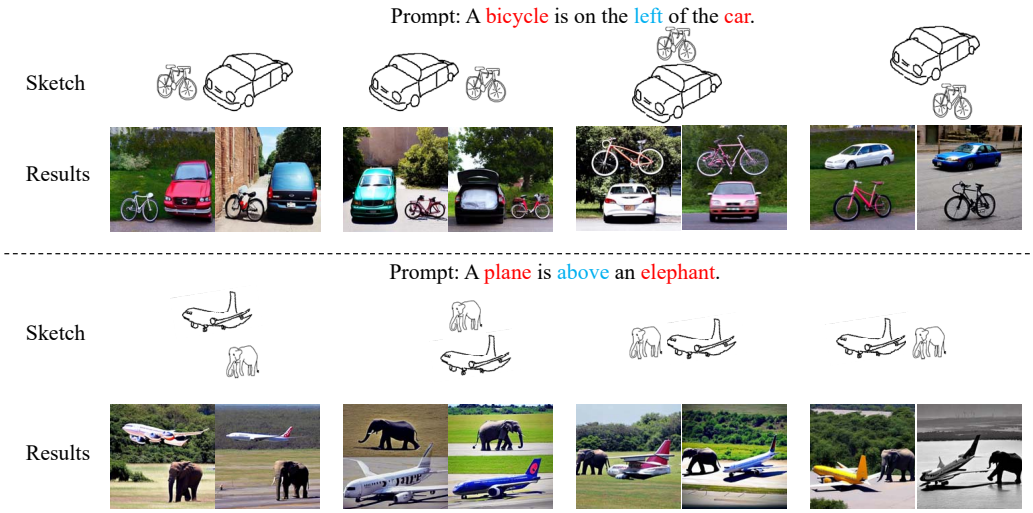


Figure 5.4: We verified that the generated image will follow the positional relationship of our sketch even if it contradicts the input text.

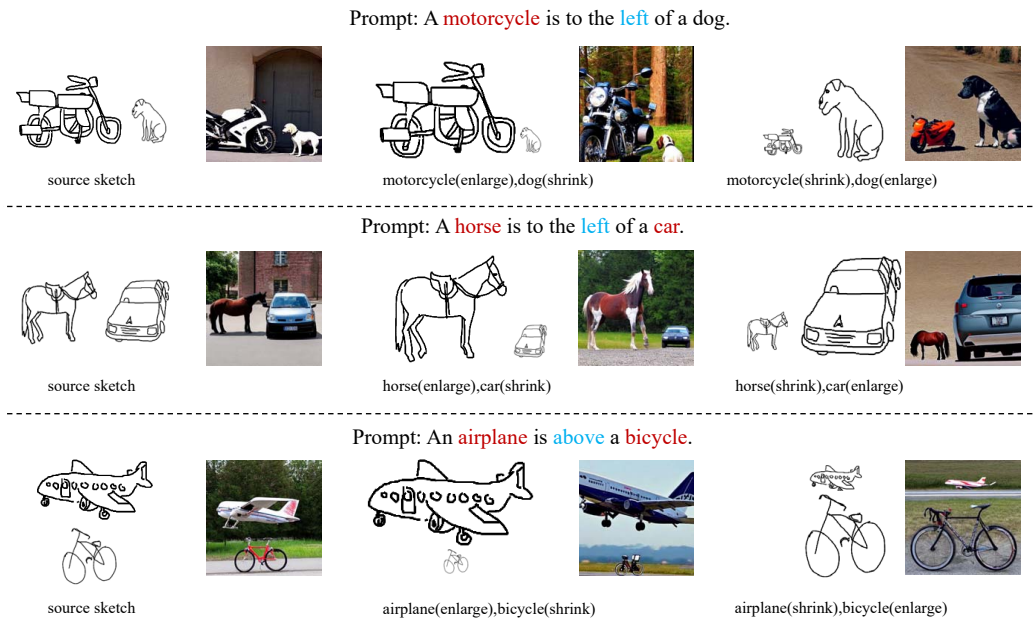


Figure 5.5: Our model can control the scale of objects in the generated image by controlling the objects' scale of the sketch.

	FID(↓)	YOLO score(mAP / AP_{50} / AP_{75})(↑)
LDM	21.42	0.5 / 2.4 / 0.4
Our model	27.34	21.6 / 42.0 / 21.7

Table 5.2: Comparison between the pre-trained LDM[1] and our model.

Model	FID(↓)
pix2pix	143.1
SketchyGAN	141.5
SketchyCOCO	87.6
Our model	**21.04**

Table 5.3: We compare the proposed method with image generation methods (pix2pix[8], SketchyGAN[6], and SketchyCOCO[7]) in image quality.

tional sketches. As shown in Figure 5.5, we change the scale of the objects in the sketches with the text prompt constant, the corresponding objects in the generated image will change accordingly, even if the generated image does not conform to realistic logic at all.

In summary, we demonstrated the model's ability to generate images with specific positions, leveraging sketches as guidance. This model helps address the issue of object misplacement in the original diffusion models. Through the object loss and scale change experiments, we have found that sketches exhibit stronger control in the attention layers compared to the original text prompts. However, it is important to note that sketches that are against common sense can lead to unnatural generated images (such as in the third example of Figure 5.2, the giraffe is bigger than the airplane weirdly).

5.3 Quantitative Comparsions

We compare our model with the state-of-the-art methods on the sketch-to-image task (pix2pix[8], SketchyGAN[6], and SketchyCOCO[7]). We also conduct a comparison study between our model with the LDM to demonstrate the usefulness of our model for position control. Since prior text-to-image methods do not support taking sketches as input, it is not fair to compare with them on this metric. Thus, we only report metrics for the LDM as a reference.

We employ Fréchet Inception Distance[69] (FID) as a metric to assess the quality of the generated images. FID quantifies the similarity between ground truth and generated samples. A lower FID value indicates that the feature

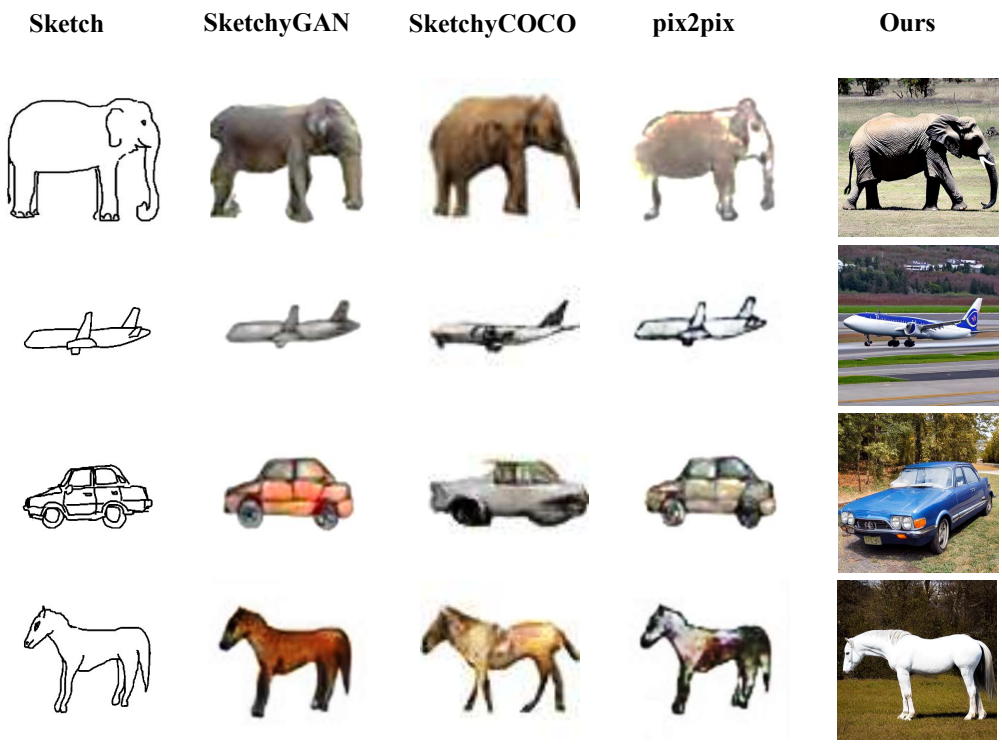


Figure 5.6: The single-object images generated by SketchyGAN[6], SketchyCOCO[7], pix2pix[8] and our proposed model. Our proposed model generates images with better quality and higher resolution than GANs.

distributions of generated samples are closer to real samples, implying better quality of the generative model. Conversely, a higher FID value suggests a larger discrepancy between the feature distributions of generated and real samples, indicating the lower quality of the generative model. To evaluate grounding accuracy (the correspondence between the input bounding box and generated entity), we use the YOLO score[70]. The YOLO score is used to indicate the confidence level of whether the target object is contained in the bounding box. It can be viewed as a score that measures the probability or confidence that a target exists. Therefore, the YOLO score can be regarded as an index to comprehensively evaluate the object detection results, which is used to measure the reliability of the bounding box and the localization accuracy of the object.

As shown in Table 5.2, our model (27.34 FID value) performs less favorably in terms of FID score compared to LDM (21.42 FID value). This is largely due to the influence of the sketches' positional control on the model's generation capability, particularly with sketches that exhibit unconventional

relative proportions. However, our model has achieved success in terms of YOLO scores (21.6, 42.0, 21.7 scores in average precision is better than 0.5, 2.4, and 0.4 scores of LDM), indicating that our model can generate corresponding objects at the desired positions.

We conduct the comparison experiment between our proposed and several image generation methods (pix2pix[8], SketchyGAN[6], and SketchyCOCO[7]) in FID value. Since the sketchyGAN and pix2pix are not models for scene images, we just utilize the single object sketches in this situation (As shown in Figure 5.6). As shown in Table 5.3, our proposed model gets the best result (21.04 FID value) in image quality than previous work, due to the strong generative ability of the diffusion model.

Chapter 6

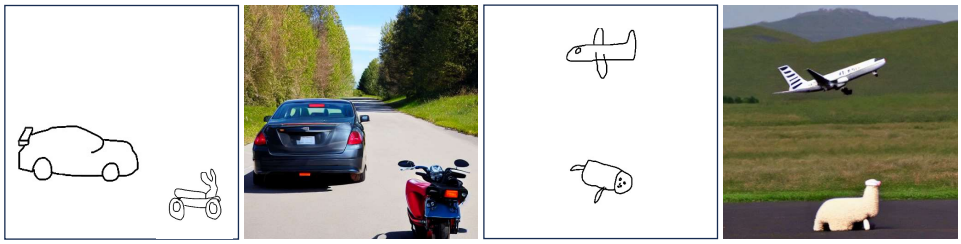
Conclusion and Limitations

6.1 Conclusion

In this thesis, we investigated the sketch-based position control by the pre-trained latent diffusion model without fine-tuning or training. Our proposed model has two stages, the feature extraction stage and the image generation stage. In the feature extraction stage, the sketches are segmented by the pre-trained segmentation model to obtain the labels and bounding boxes. The labels and bounding boxes will be encoded as grounding tokens, which are injected into cross-attention layers to guide the position for image generation. In the image generation stage, the model use pre-trained LDM to generate images. The constant α divides the diffusion process into two steps. We explored that different α will affect the generative ability of LDM and positional control of generated images.

Our method can obtain the generated image, whose objects' positions are consistent with the sketches', and effectively solve the object loss issue of the original LDM. We also explored the control conditioning priority of the attention layers will be greater than the original text prompt. Sketches as conditional input can precisely control the relative position between objects, even if this position does not match the text description. In addition, sketches have a significant effect on controlling the scale of the object. We change the position and scale of the objects of sketches in the experiment, the corresponding position and scale of the objects in generated images will change accordingly, even if the content described by sketches is completely contrary to text prompts. By controlling the position and relative scale of the sketches, we can precisely control the position and scale of the generated objects, which is difficult to achieve in the original text prompts.

Moreover, we attempted to explore the ability to generate freehand sketches



(a) (b)

Figure 6.1: Free-hand sketch generation results that do not belong to the SketchyCOCO dataset with text prompts (a): “a car and a motorcycle” and (b): “an airplane and a cat”.

beyond the SketchyCOCO dataset. As shown in Figure 6.1,(a) demonstrates a successful case where the model accurately predicted the objects in the sketch and guided the generation of the image's position. However, (b) illustrates a failure case where due to limitations in drawing ability, the intended “cat” was misidentified as a “sheep”, resulting in conflicting conditional inputs and text prompts. Despite the image generated based on conditional input, the output appears unnatural. Overall, free-hand sketches reflect human creativity and artistry, allowing the proposed models to be widely applicable to each user. However, due to differences in drawing skills and experience, the quality of free-hand sketches varies, posing challenges in tasks such as sketch segmentation and recognition.

We also conduct quantitative comparisons with LDM and GANs. The results show that our proposed model achieves position control and gets better image quality than GANs.

6.2 Limitations and Future Work

Our model still has many limitations. First, as shown in Figure 6.2, sketches contain not only position information but also shape and details. Sketch lines can represent an object's boundaries and shapes. Major features can be presented by shading cues or some simple color markers to convey information about lighting, shading, and color. As shown in the first row of Figure 6.3, our current model is only limited to using the sketch to control position information and does not make full use of all the advantages of sketches. The “elephant” in (a) faces left in the sketch but faces right in the generated image. Similarly, the “horses” in (b) are facing opposite directions, and the “horse” in the image is even incomplete. In (c), the “cloud” and “airplane”

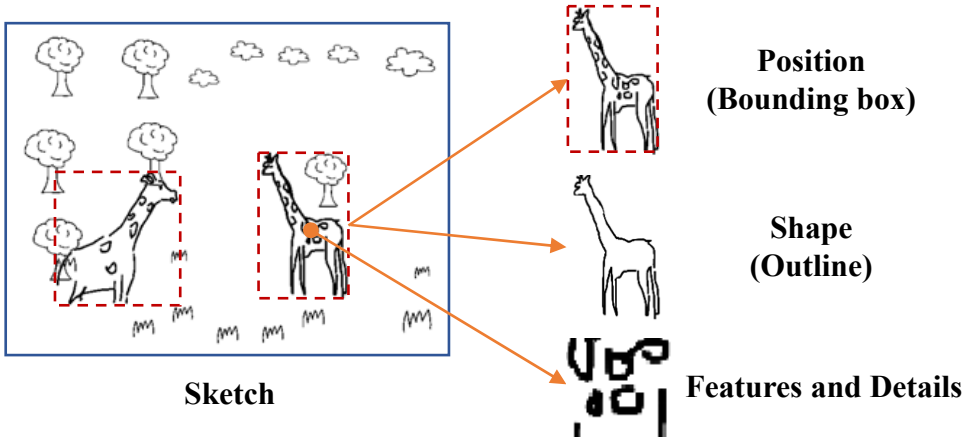


Figure 6.2: Sketch provides the structure and composition of the target image, encompassing outlines, shapes, and detailed information.

are not match the shape in the sketch, even though the positions are correct. In addition to bounding boxes and labels, we can extract edge lines and masks that express detailed shape contours of sketches. Based on the detailed contour lines, we can represent them in attention maps. For example, previous work[3] represented pixels within the contour as constant $C > 0$ and pixels outside the contour as 0, thus visualizing the contour on the attention map.

Previous work guides a diffusion model with a spatial map to enable pixel-level controlled generation of individual sketch objects[52]. However, The method encounters difficulties when confronted with complex and cluttered sketches because it treats all strokes uniformly, without giving priority to their saliency or semantics. Combining the pixel-by-pixel sketch-to-image method with our model and thus extending it to scene sketches is another direction for our future work. We plan to perform object segmentation in scene sketches and generate individual sketch objects at the pixel level. Subsequently, we will apply the object-to-image fuser module from PasteGAN[71] to merge the generated objects into a single image.

Moreover, the proposed model utilizes the sketch as additional conditional information, the semantics and intentions of sketches may not be unambiguous and limited by the user's drawing level. As shown in (d) of Figure 6.3, the left animal is a "dog" according to the text prompt, but the rough sketch was identified as "horse".

In addition, using other conditional control information is also a common operation in image generation. How to compare our model with other existing

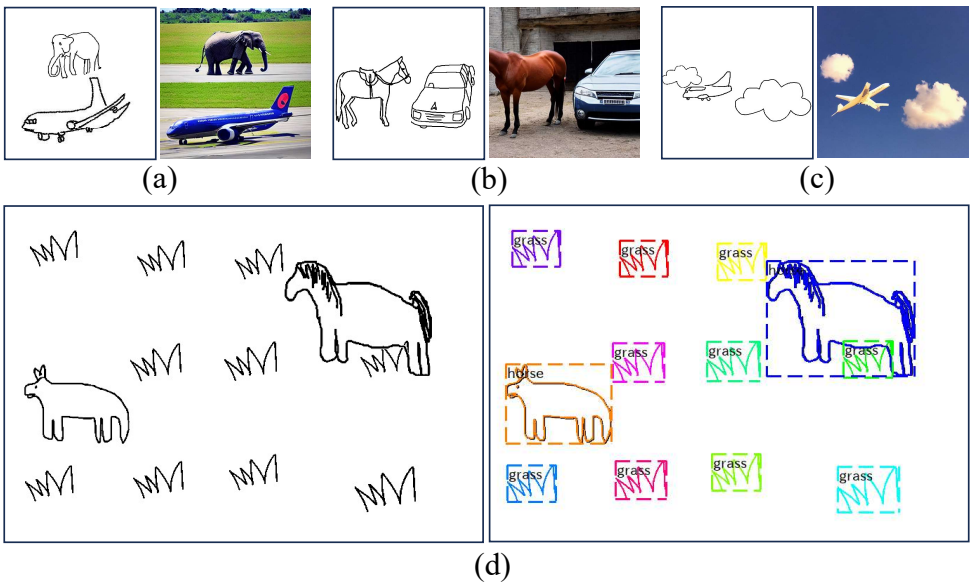


Figure 6.3: Some failure cases of the proposed model. (a), (b), and (c) indicate that the proposed model only performs object position guidance, but it cannot achieve shape or pixel-level control. (d) shows the failure segmentation case.

models and combine it with a more multi-modal input conditional control model is also one of the more novel works in the future.

Acknowledgement

I would like to give my heartfelt thanks to all the people who have ever helped me with this thesis.

My sincere and hearty thanks and appreciations go firstly to my supervisor, Prof. Haoran Xie, whose patience, and immense knowledge give me significant guidance and help. Prof. Xie patiently explained the deficiencies in my thesis and put forward valuable comments for improvement. Prof. Xie also taught me a rigorous research attitude, that is, each part must be clearly understood and supported by evidence, even a small symbol definition. It has been a great privilege to study under his guidance and supervision.

Additionally, I would like to express my heartfelt gratitude to Professor Miyata and Professor Yoshitaka for their dedicated guidance and professional advice in my academic research. Professor Miyata's patient guidance and encouragement have provided me with endless inspiration and support in my research, helping me adapt to living in Japan for the first time. Moreover, Professor Miyata's assistance in Japanese writing has been of immense help. I am also deeply thankful to Professor Yoshitaka for his meticulous guidance and valuable suggestions on my major research topic. His comments during the initial proposal stage refined and clarified my ideas, making my thesis more comprehensive and concrete.

Furthermore, I would like to express my gratitude to all the members of our lab for their continuous support from the initial stages to the completion of this thesis. During our lab meetings, they attentively listened to my research ideas and provided valuable insights and suggestions based on their practical experience. I am especially thankful to Dr. Du Xusheng for his assistance with the thesis's formatting and various aspects of my life. His unwavering dedication during my illness, making multiple trips to the hospital to help me, greatly alleviated the obstacles I faced in completing the thesis.

In addition, many thanks go to my family for their unfailing love and unwavering support.

References

[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] 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.

[3] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," *arXiv preprint arXiv:2304.03373*, 2023.

[4] 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.

[5] 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.

[6] W. Chen and J. Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9416–9425.

[7] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5174–5183.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[9] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

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

[11] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 2022, pp. 88–105.

[12] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[13] Z. Huang, Y. Peng, T. Hibino, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "dualface: Two-stage drawing guidance for freehand portrait sketching," *Computational Visual Media*, vol. 8, pp. 63–77, 2022.

[14] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "Difffacesketch: High-fidelity face image synthesis with sketch-guided latent diffusion model," *arXiv preprint arXiv:2302.06908*, 2023.

[15] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*. Springer, 2022, pp. 17–33.

[16] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 062–18 071.

[17] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 1033–1038.

[18] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International conference on machine learning*. PMLR, 2015, pp. 1462–1471.

[19] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[20] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei, "Avoiding latent variable collapse with generative skip models," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2397–2405.

[21] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[23] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 743–10 752.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[26] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[27] 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.

[28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics,"

in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[30] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[31] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *International Conference on Learning Representations*, 2021.

[32] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[33] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 981–17 993, 2021.

[34] 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.

[35] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 49–58.

[36] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[37] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 2022, pp. 89–106.

[38] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[39] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[40] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[41] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[42] 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.

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

[44] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[45] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.

[46] J. S. Ubhi, A. K. Aggarwal *et al.*, "Neural style transfer for image within images and conditional gans for destylization," *Journal of Visual Communication and Image Representation*, vol. 85, p. 103483, 2022.

[47] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5524–5532.

[48] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong gan: Continual learning for conditional image generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2759–2768.

[49] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[50] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," *arXiv preprint arXiv:2301.07093*, 2023.

[51] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual gan," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 205–220.

[52] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," *arXiv preprint arXiv:2211.13752*, 2022.

[53] J. Weng, X. Du, and H. Xie, "Dualslide: Global-to-local sketching interface for slide content and layout design," *arXiv preprint arXiv:2304.12506*, 2023.

[54] Y. Peng, C. Zhao, H. Xie, T. Fukusato, K. Miyata, and T. Igarashi, "Dualmotion: Global-to-local casual motion design for character animations," *IEICE TRANSACTIONS on Information and Systems*, vol. 106, no. 4, pp. 459–468, 2023.

[55] Z. Huang, H. Xie, T. Fukusato, and K. Miyata, "Anifacedrawing: Anime portrait exploration during your sketching," *arXiv preprint arXiv:2306.07476*, 2023.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[58] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Effective conditioned and composed image retrieval combining clip-based features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 466–21 474.

[59] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[60] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchyscene: Richly-annotated scene sketches," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 421–436.

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

[62] 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.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[64] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[65] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[66] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.

[67] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1209–1218.

[68] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[70] Z. Li, J. Wu, I. Koh, Y. Tang, and L. Sun, "Image synthesis from layout with locality-aware mask adaption," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 819–13 828.

[71] Y. Li, T. Ma, Y. Bai, N. Duan, S. Wei, and X. Wang, "Pastegan: A semi-parametric method to generate image from scene graph," *Advances in Neural Information Processing Systems*, vol. 32, 2019.