Master's Thesis

Sketch-Guided Consistent Image Generation Using Multi-Aggregation Attention Mechanism

YANG SHIHAO

Supervisor HAORAN XIE

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

December 2024

Abstract

Current image generation methods involve users inputting textual information as prompt, which serves as a text-based conditional control input to guide the image generation process. This text-based approach may have difficulty meet the user's requirements and initial design intentions in the generated image. In common cases, the users shall either regenerate the image from scratch or manually modify the generated image using their drawing skills. These design processes are normally time-consuming and labor-intensive, and it requires a certain level of drawing skills and experiences from the users. Recently, deep learning-based image generation models have demonstrated remarkable success in generating high quality images. Generating coherent image sequences is essential for applications such as manga, animation, and other visual storytelling mediums, where maintaining consistency across multiple frames is critical to preserving narrative flow and visual continuity. However, the state-of-the-art image generation current models still face challenges in generating coherent image sequences, such as insufficient consistency in image features, including character appearance, background elements, and object positions across frames. Generating high-quality and consistent image sequences is crucial for storytelling and visual effects in video game development, virtual reality experiences, and film production. Traditional image generation methods often fail to meet the requirements of maintaining feature consistency across sequential images, necessitating the development of new methods to enhance image generation quality and consistency.

This thesis proposes a two-stage image generation model designed to generate multiple sequential images based on a series of coherent text prompts. In the first stage, the model generates a set of sketches based on the user's textual input. Users can modify these sketches and arrange them according to their design intentions. In the second stage, the model utilizes the revised sketches from first stage to generate a series of images with consistent image features. In this work, we propose the Multi-Aggregation Attention (MAGA) mechanism in the second stage, which can significantly enhance feature aggregation and refinement, thereby improving image consistency. The MAGA mechanism incorporates an external memory component during the image generation process, storing conditioning information from previous network layers and establishing connections between the current text, sketches, and spatial features, and all historical inputs including prior

text and sketch information, thereby effectively increasing consistency across sequential images.

To enhance user flexibility, we proposed a sketch selection workflow, offering two more options. First, users can manually input their own sketches, which allows for highly customized and precise visual elements that reflect specific user intentions. Alternatively, users can select sketches from a high-quality, pre-constructed database designed to support the image generation process. This sketch database, called SketchXL, includes about 1,500 coherent storylines of continuous sketches, thus addressing the lack of sequential consistency found in existing datasets.

Through a series of quantitative and qualitative experiments, we verified that our model can demonstrate outstanding performance in image quality, consistency, and fidelity of the generated images from the input sketches and text prompts. Firstly, the proposed two-stage model demonstrated an advantage over single-stage generation models by allowing users to adjust the final output by editing the sketches generated in Stage 1, rather than the final generated images. This approach can not only simplify the drawing process but also enhance the quality of the generated images. In our experiments, we compared the proposed model against several state-of-the-art models under the same input conditions. We evaluated the generated images for consistency and detail preservation to the input sketches and text prompts. The experimental results show that the proposed model has significant advantages in generating continuous image sequences, maintaining consistency in character features and background images while preserving details. Additionally, we quantified image quality using the NIMA score, with the proposed model achieving an overall image quality score of 6.04, outperforming other state-of-the-art models (5.84 for Stable Diffusion v1.0, 5.99 for Stable Diffusion XL, 4.95 for ControlNet). Overall, our method provides an efficient and high-quality image generation solution for the applications of animation and comic creations, with broad potential applications.

Keywords: Latent Diffusion Model, Attention Mechanism, Consistent Images Generation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image generation represents a significant research direction in the fields of computer graphics, artificial intelligence and computer vision [3] [1]. The primary goals of image generation include enhancing the quality, diversity, and controllability of generated images while exploring broader application domains. Current applications encompass areas such as facial recognition [4] [5] [6], image reconstruction [7] [8], style transfer [9] [10], and the creation of artistic works [11] [12]. In addition, image generation has become increasingly important in industry applications, including anime and manga production [13] [14], fashion design for generating creative concepts [15] [16], and advertising, where high-quality and visually appealing images are needed to engage consumers. These applications highlight the growing impact of image generation technologies in both creative and commercial domains.

Currently, diffusion models have emerged as cutting-edge models in the field of image generation, achieving state-of-the-art performance in various areas such as image generation and multimodal conditional control. Prominent large models that utilize diffusion models as their backbone, including Stable Diffusion and DALL-E [2], not only lead research directions in the scientific community but have also gained widespread recognition among AI enthusiasts and artists [17] [18]. The extensive user community has fostered the development of diverse pre-trained diffusion models and the creation of a vast array of artistic works using these models.

Sequential images are a series of interconnected visual elements that maintain coherence in terms of style, content, and structure, much like the four-panel comic strip as shown in Figure 1.2. Each panel contributes to the progression of the narrative, with each subsequent image building upon the previous one to advance the storyline. The consistency of visual elements such as character appearance, expressions, backgrounds, and stylistic features plays a crucial role in maintaining the flow of the story. Sequential image generation is a significant research direction in the fields of computer vision and image generation, aiming to produce coherent sequences of images. These image sequences should maintain consistency in content, style, and structure, and are widely applied in animation production [19], video generation [20],

Figure 1.1: The comparison of image generation results from different models when processing consistent text prompts. The proposed model that introduced in this paper performs the best in terms of consistency. The visual discrepancies between images generated by this model are minimal, indicating that it better maintains image consistency and feature continuity when generating images from sequential text prompts. In contrast, Stable Diffusion 1.0 and Stable Diffusion XL exhibit poorer consistency, while DALL-E 3 and Stable Diffusion 1 with ControlNet perform at an intermediate level.

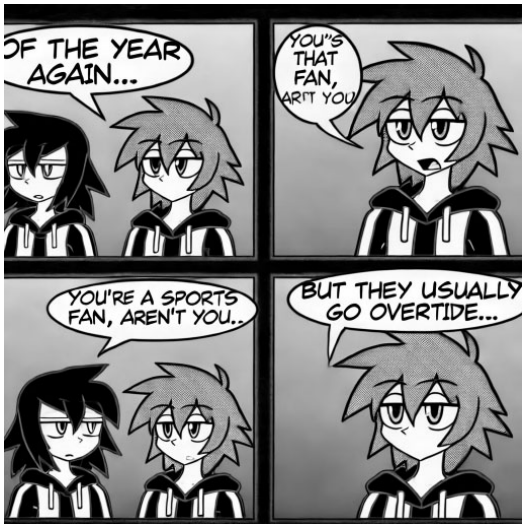


Figure 1.2: Sequential images generated by our method. Image consists of a four-panel comic strip featuring two anime-style characters engaging in a conversation. Across the panels, the characters maintain consistent features, such as their hairstyles and clothing.

and storyboard design [21].

Compared to generation of single image, consistent image generation imposes higher demands on models regarding temporal consistency and feature preservation. Image generation models need to meet a series of specific consistency requirements. Consistency is one of the core requirements in this domain. Models need to consider the temporal dependencies between consecutive images to ensure that the entire image sequence remains coherent in terms of content. Additionally, image generation models should preserve key features consistently throughout the generation process, such as character features, poses, and background elements of characters. Lastly, the generated images must maintain high quality and diversity to accommodate various application scenarios, such as animation production, manga and comics, and video game cutscenes etc.

Despite significant advancements in single-image generation tasks, current image generation models exhibit several shortcomings when generating consistent image sequences, failing to fully meet the demands of this task [22]. Present image generation models primarily focus on producing high-quality individual images rather than consistent sequences [23] [24]. Although they perform well in generating single images, they lack the necessary mechanisms to ensure consistency across successive images [25]. This deficiency results in issues such as feature drift and inconsistencies in character attributes, poses,

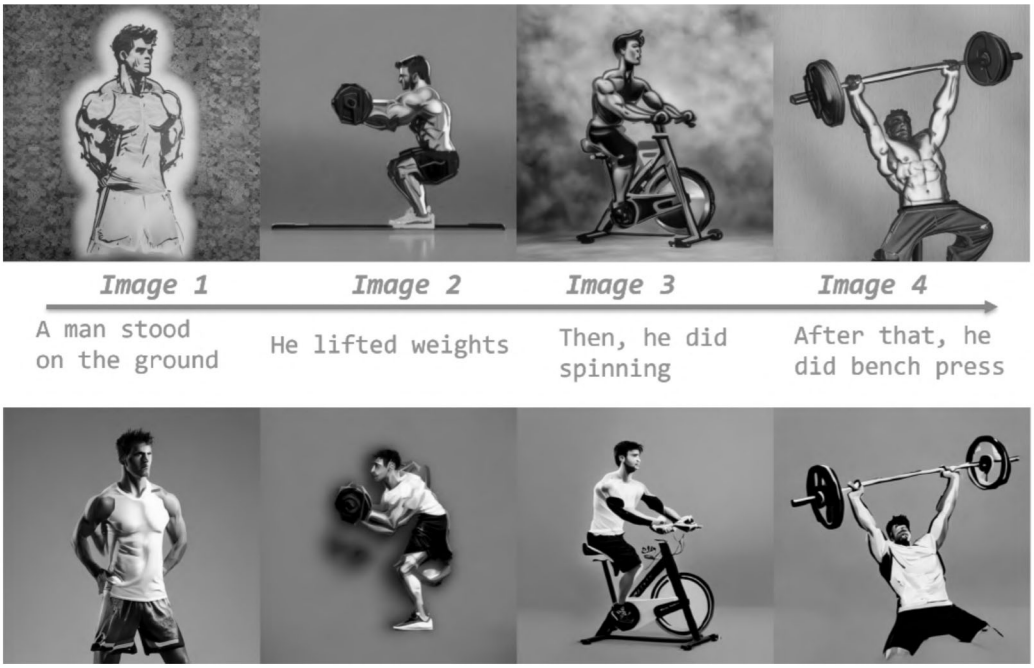


Figure 1.3: The comparative results show two rows of images: the upper row illustrates images generated by Stable Diffusion 1.0 with ControlNet, and the lower row showcases the output from our proposed model. The upper row highlights inconsistencies in the character's appearance and posture across different sports activities, while the lower row demonstrates how our model maintains consistent character features and postures, ensuring greater continuity and stability throughout the sequence.

and background elements between consecutive images.

For example, as shown in Figure 1.3, when using the Stable Diffusion 1 with ControlNet [26] (Controlnet), the character's features vary significantly between different prompts, failing to maintain a cohesive visual identity. Additionally, each activity introduces new postures and equipment, but the transitions are not smooth, and the changes are too abrupt, further breaking the continuity. The overall style, including background and character details, fluctuates across images, undermining the stability and consistency of the character depiction.

In summary, current sequential image generation task exhibits several shortcomings, include: 1) The process of generating images directly from textual information often results in errors in the generated images, and these images are challenging for users to modify. These errors typically involve inaccuracies in spatial relationships, character actions, or object positioning,

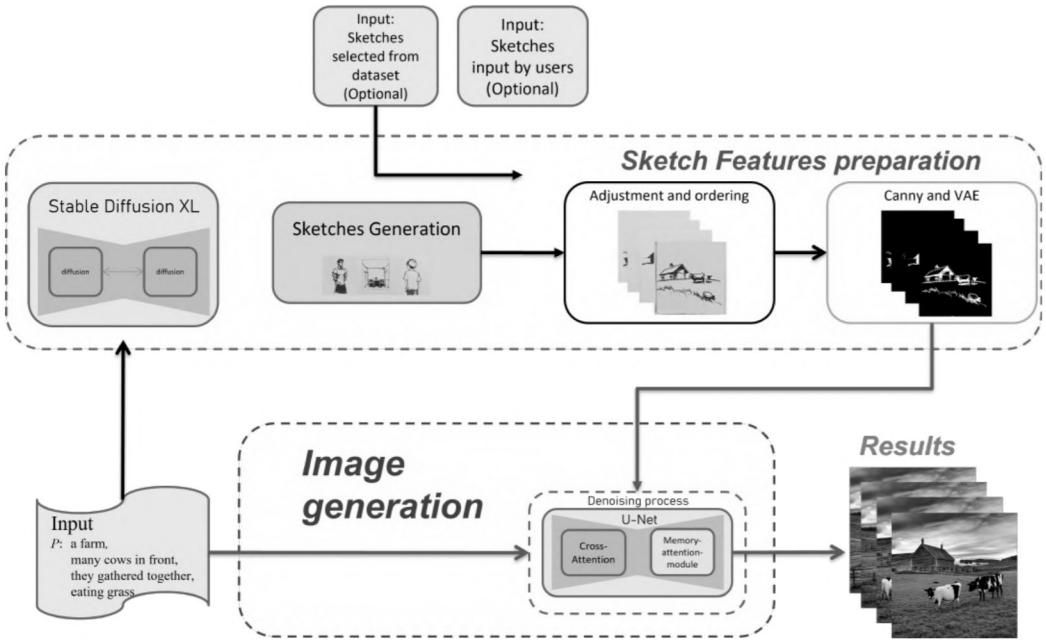


Figure 1.4: The proposed pipeline includes the preparation of sketches, the generation process, and the reasoning result. Since the model is a tuning-free method, no more training is necessitated.

where the generated images fail to reflect the intended descriptions from the text prompts. 2) Existing models fail to ensure temporal consistency and feature preservation across multiple frames, which are crucial for creating sequential images. These results in issues like feature drift and inconsistencies in character attributes, poses, and background elements between consecutive frames.

To address the aforementioned challenges, this paper proposes a two-stage architecture for sequential image generation. This method involves several key stages and innovations aimed at enabling users to modify the final output images during the generation process and overcoming the deficiencies of existing models in maintaining temporal consistency and feature preservation.

Our approach divides the text-to-image generation process into two distinct stages. In the first stage, sketches are generated based on the user's text prompt, allowing users the flexibility to adjust and arrange these sketches.

This research also provides a sketch dataset for users to select and input into the network for the second stage. Additionally, for users with advanced drawing skills, the system permits the upload of their own sketches. Users can then arrange these sketches in the desired sequence, laying the groundwork

for their image narrative.

In the second stage, the sketches and textual prompts are encoded and used as conditions in our proposed sequential images generation model. This model focuses on maintaining temporal consistency and feature preservation across multiple images, ensuring that character features, poses, and background elements remain consistent throughout the generated sequence.

This thesis includes the following main contributions:

- First, we introduce a two-stage image generation process. In the first stage, corresponding sketches are generated based on textual descriptions, allowing users to modify, adjust, and arrange these sketches. In the second stage, both the sketches and the text are used as conditional inputs to the proposed model, guiding the generation process to produce a coherent sequence of images.
- Second, we propose a new model specifically designed for consistent image generation. This model is capable of generating multiple images while maintaining consistency in character features, poses, and background elements. By addressing the key limitations of existing models, our approach ensures that the generated images are coherent and visually consistent, even across multiple frames.

Chapter 2

Related works

2.1 Image Generation

In recent years, diffusion models have become a focal point of research in image generation field because of their robustness, stability during training, and ability to produce high-quality images. Unlike Generative Adversarial Networks [27] (GANs), which often suffer from training instability and mode collapse, diffusion models offer a more stable and reliable approach [3]. Their iterative denoising process allows for fine-grained control over the image generation, leading to superior results in terms of both fidelity and diversity.

During the early stages of image generation development, most methods relied on manual feature extraction or statistical models. Researchers initially modeled the input images and then utilized these models to generate new images [28] [29]. However, with the advancement of deep learning, deep learning based image generation models have demonstrated significant advantages. Among these advancements, Generative Adversarial Networks (GANs) have particularly stood out, representing a major breakthrough in image generation technology.

GANs consist of a generator and a discriminator. During the training process, the generator is responsible for producing images that closely resemble those in the training set, while the discriminator's role is to distinguish whether an image originates from the dataset or is generated by the generator. This adversarial training process ultimately enables the generator to produce images with a distribution that closely approximates that of real images.

Additionally, conditional generation models allow users to incorporate conditions to enhance control over the image generation process. [30] [31]The conditions that users can provide include textual information, hand-drawn sketches, etc.

Despite the significant success of GANs in the field of image generation, several critical drawbacks pose considerable challenges for practical applications. The adversarial training process between the generator and

the discriminator can lead to non-convergence issues [32]. Additionally, mode collapse is a common and severe problem in GAN training, where the generator only learns a limited subset of the data distribution [33]. This limitation prevents GANs from generating images that fully reflect the diversity of the real data distribution, resulting in a lack of variety in the generated images.

In recent years, diffusion models have rapidly advanced, making significant contributions to computer vision and image generation field. diffusion models are known for generating high-quality images with substantial diversity. Denoising Diffusion Probabilistic Models (DDPM) are generative models based on a gradual denoising process [19]. The core idea of DDPM is to simulate the image generation process by progressively denoising to a clear image. This generative process can be divided into two stages: the forward diffusion process and the reverse denoising process. During the forward diffusion process, noise is gradually added to the input image until it becomes pure noise. This process is achieved through a series of steps, each adding noise according to a predefined noise distribution. In the reverse denoising process, the model starts from random noise and generates an image by progressively removing the noise. Each denoising step is performed by a deep neural network that learns to reverse the forward process. However, DDPM requires a large number of denoising steps to generate high-quality images. Each step in this process involves a detailed transformation, which cumulatively results in longer generation times and higher computational costs. Therefore, aiming to further enhance generation efficiency and image quality, Denoising Diffusion Implicit Models (DDIM) significantly reduces the time required for image generation through an improved sampling process [34]. Latent Diffusion Models are designed to perform diffusion processes in the latent space of a pre-trained autoencoder, rather than directly on the high-dimensional pixel space [17]. This approach leverages the dimensionality reduction and feature extraction capabilities of autoencoders, resulting in more efficient and scalable image generation than DDIM.

Diffusion models operate based on a two-phase process: a forward diffusion process that adds noise to the data, and a inference process that denoises the data to generate new samples [35] [36].

As shown in Figure 2.1 ,The forward diffusion process is a method used to gradually add noise to an initial data sample over time steps, transforming it into pure noise by the final step. Imagine starting with a clear photograph, and progressively adding layers of static noise until the image is fully obscured. This transformation happens through incremental steps, with each step adding a small amount of noise, making the image less recognizable as it progresses.

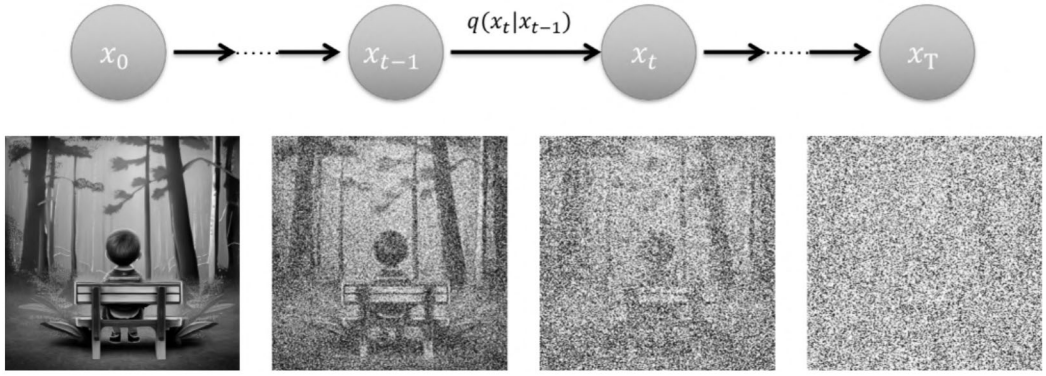


Figure 2.1: The forward diffusion process of a diffusion model

For a data point, sampled from the real data distribution q(x), the forward diffusion process is modeled by adding Gaussian noise at each step of a Markov chain. Specifically, at each step t, Gaussian noise with variance is added to the previous state, generating a new latent variable with the distribution. This process is illustrated as the gradual transition of an image from its original state to an entirely noisy representation.

The inference process aims to recover the original data from the noisy sample produced at the end of the forward process. This is achieved by learning a model that can reverse the noise addition, step by step. The inference process is like peeling away the layers of noise that were added during the forward process. Starting from the highly noisy image, the model iteratively removes noise, gradually refining the image back towards its original state.

To train the model, the process involves guiding it to accurately predict the noise that was added at each step of the diffusion process. By learning to reverse the effects of this added noise, the model gradually becomes more adept at reconstructing the original image. This training is conducted using a large dataset where both the original images and their noisy versions are provided, allowing the model to improve as it encounters more examples.

Once the model is trained, generating a new image involves starting with a random noise image and applying the learned denoising steps in reverse order. This iterative process starts with a completely noisy image and, through successive steps, refines it into a coherent and high-quality image. Each step slightly reduces the noise, progressively revealing more details of the final image. By the end of the process, the model produces an image that closely resembles the type of images it was trained on.

Diffusion models laid the groundwork for understanding and developing more advanced and conditional variants. These models highlighted the

potential of using iterative refinement and probabilistic modeling to achieve high-quality image generation, establishing a robust foundation for future research in the field. However, diffusion models also possess several limitations. First, the iterative nature of the denoising process means that generating a single image can be computationally expensive and time-consuming. Each step in the process requires significant computation, making it less efficient compared to some other image generation methods. Second, due to the numerous steps required in both the forward and reverse processes, the time taken to generate an image is relatively long. This slow sampling speed can be a bottleneck in applications requiring real-time or near-real-time image generation. Most importantly, early diffusion models do not incorporate conditional inputs, limiting their ability to generate images based on specific conditions or attributes. This restricts their applicability in scenarios where controlled and specific image generation is required.

This section has outlined the key principles of diffusion models in image generation, emphasizing their strengths and limitations. While diffusion models offer high-quality image generation through iterative denoising, they suffer from computational inefficiency and slow sampling speeds due to the multiple steps involved. Furthermore, early diffusion models lack conditional control, limiting their applicability in scenarios that require specific, user-defined outputs.

This paper builds upon these foundational models by introducing a two-step generation process that integrates the Multi-Aggregation Attention (MAGA) mechanism. This innovation not only improves consistency across sequential images but also introduces conditional inputs, allowing for greater user control. In doing so, our work addresses the existing limitations of diffusion models and extends their applicability to generating coherent and sequential images efficiently.

2.2 Conditional Image Generation

Conditional image generation refers to the process of generating images based on specific conditions or inputs. These conditions can take various forms, such as sketches, textual descriptions, or other images. By incorporating these conditions, models can produce images that align closely with the given inputs, providing a higher degree of control over the generated content. Conditional image generation provides a wide aspects of applications, including Medical Imaging [37] [38], style transfer [39] [40], and real image editing [41] [42].

Traditional conditional image generation primarily relies on Generative

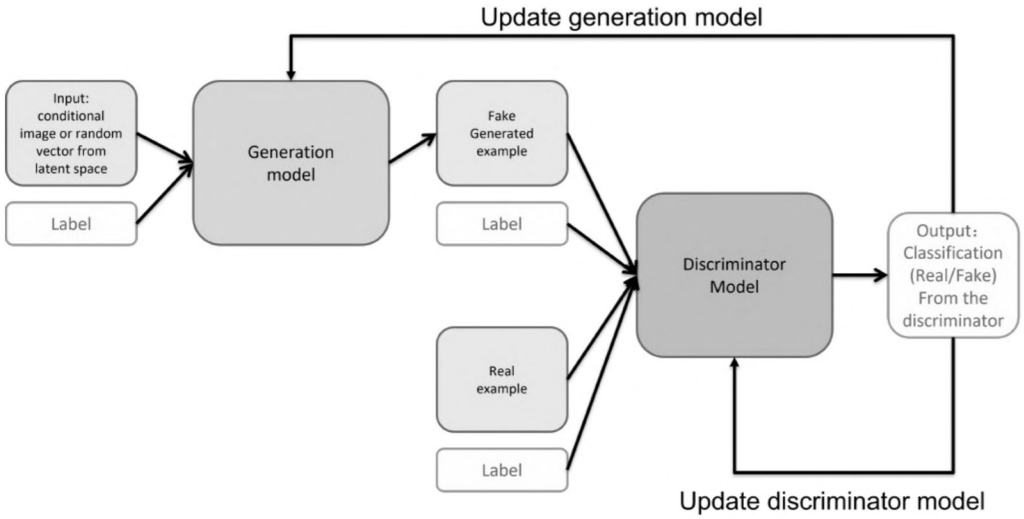


Figure 2.2: Architecture of Conditional Generative Adversarial Networks

Adversarial Networks [27] (GANs). Conditional GANs [31] (cGANs) is one of the most influential works in conditional image generation. This model adapts the basic GAN framework to condition the generation process on auxiliary information, enabling a wide range of image-to-image translation tasks.

Conditional Generative Adversarial Networks (cGANs), shown in Figure 2.2 are a type of GAN. The generator takes both the input image and the condition as inputs and generates an output image that aims to satisfy the given condition. The discriminator receives both the input condition and the generated (or real) image and predicts whether the image is real (from the training dataset) or fake (from the generator). The architecture typically includes convolutional layers, batch normalization, and activation functions, focusing on distinguishing real images from fake ones conditioned on the input.

The cGAN framework [30] has been applied to various image-to-image translation tasks, demonstrating its versatility and effectiveness. Some notable applications include sketch-to-photo translation, style transfer and super-resolution. These applications highlight the model's ability to handle diverse tasks by leveraging the conditional information provided during generation.

Recent developments in the field of conditional image generation have seen a shift from GAN-based approaches to diffusion models. Unlike GANs, diffusion models, on the other hand, offer a more stable and robust framework

for image generation. They provide finer control over the generation process through iterative denoising, leading to superior image quality.

Chen proposes a novel method that combines reference images and sketches using a structure-aware diffusion model [43]. The key innovation of this work lies in its ability to effectively integrate structural information from sketches and stylistic details from reference images. The structure-aware diffusion model ensures that the generated images align closely with the provided sketches while maintaining the style and texture of the reference images. This dual conditioning mechanism allows for the creation of highly detailed and contextually appropriate images, addressing the common challenge of balancing structural integrity with stylistic coherence.

Zhang et al. introduces a framework called ControlNet that enhances pre-trained diffusion models with additional conditional controls [26]. The primary innovation of ControlNet is its ability to integrate diverse conditional inputs, such as edge maps, depth maps, or textual descriptions, into the image generation process. By adding a conditional control branch to the pre-trained diffusion model, ControlNet allows for more precise and flexible generation of images based on multiple conditions. This capability significantly extends the utility of diffusion models, enabling them to handle complex image generation tasks that require adherence to multiple constraints. The authors highlight the versatility of ControlNet in various applications, including guided image synthesis and enhanced image editing, demonstrating its effectiveness in generating contextually appropriate and high-quality images.

Compared to traditional unconditional image generation, conditional image generation offers the significant benefit of guiding the generation process using specific inputs, such as text descriptions, sketches, or reference images. This results in outputs that are more relevant and aligned with user intentions. However, the introduction of conditions also brings about greater challenges. One of the significant challenges is ensuring that the generated images faithfully adhere to the provided conditions while maintaining high quality. Additionally, conditional models often face increased computational complexity and require extensive training data that covers a wide range of conditions. Ensuring coherence in the generated images, especially when multiple conditions are involved, is another critical challenge that researchers must address.

2.3 Sketch-based Image Generation

A sketch is a simplified, often monochromatic drawing that captures the essential contours and shapes of an object or scene. Therefore, using sketches

as conditions for image generation offers several distinct advantages. Sketches provide a clear and precise representation of the structural layout of the desired image. This structural clarity makes it easier for the generative model to understand the fundamental shapes and spatial relationships within the scene, leading to more accurate and coherent outputs.

Auto-painter propose a method for generating cartoon images from sketches using cGANs [44]. In this context, the generator produces images based on a given sketch, while the discriminator evaluates the correspondence between the generated image and the input condition, ensuring the output aligns with the provided sketch [30]. The key innovation in this work lies in its ability to transform simple sketches into fully colored and detailed cartoon images. The notable strengths of this approach are its ability to produce visually appealing and stylistically consistent cartoon images that closely adhere to the provided sketches.

Sketch2Color also explores the use of cGANs to generate color images from sketches [45]. By training the cGAN on paired datasets of sketches and corresponding color images, the model learns to translate the monochromatic sketches into richly colored images, maintaining the original structure and adding realistic textures. However, while the model excels at generating detailed color images from clear sketches, it may not perform as well with complex scenes or intricate details that are not well-defined in the sketches.

SketchyCOCO [46], raised by Gao et al, presents an innovative approach to generating images from freehand sketches using a large-scale dataset. One of the primary innovations is the creation and utilization of the SketchyCOCO dataset, which contains paired data of freehand sketches and corresponding images. This dataset is specifically designed to address the variability and ambiguity inherent in freehand sketches, providing a rich source of training data for the model. The introduced model, which is called EdgeGAN, have the ability to handle the inherent noise and imprecision of freehand sketches. The authors employ advanced preprocessing techniques and robust training methodologies to ensure that the model can accurately interpret and generate images from sketches that may vary widely in quality and detail.

With the development of conditional diffusion models, sketch has been widely used for guiding the inference process of diffusion models. Andrey et al first combines sketches with textual descriptions to guide the image generation process using diffusion models [47]. By using sketches, the model gains explicit structural information, while textual descriptions provide additional context and details that enhance the richness of the generated images. However, integrating dual inputs increases the complexity of the model, which can result in higher computational requirements and longer

training times.

DiffSketching focuses on providing users with the ability to control the image synthesis process through sketches, enhancing the precision and relevance of the generated images [48]. DiffSketching allows users to manipulate the structure and content of the generated images through sketches. This mechanism ensures that the generated images closely follow the outlines and structural hints provided by the sketches, offering a high degree of control over the final output.

It is worth noting that all the models mentioned above may require high-quality sketches with clear and precise lines to generate accurate and coherent images [47] [48] [44] [45] [46]. The quality and clarity of the input sketch directly impact the output, as ambiguous, incomplete, or poorly drawn sketches can lead to suboptimal results. This requirement limits the accessibility and usability of these models for a broader range of users, particularly those without advanced sketching skills.

The studies presented in this section discuss various approaches to sketch-conditioned image generation, primarily using Conditional Generative Adversarial Networks (cGANs) and diffusion models. These works highlight the effectiveness of sketches in providing structural information for image generation and demonstrate that models such as Auto-painter, Sketch2Color, and SketchyCOCO can transform simple sketches into detailed and coherent images. However, many of these approaches require high-quality sketches with precise outlines to produce satisfactory results, limiting accessibility for users without advanced sketching skills.

This section explores sketch-guided image generation. The method proposed in this thesis builds upon these existing methods but seeks to address the limitations of requiring high-quality sketches. By introducing a novel two-step image generation framework that allows users to modify and refine sketches during the generation process, our method offers greater flexibility and accessibility. This approach not only ensures that the generated images align more closely with user intent but also lowers the barrier for users who may not have professional sketching abilities. Thus, the research discussed in this section provides foundational insights, while our work aims to enhance and expand upon these concepts to offer a more user-friendly and flexible solution for sketch-based image generation.

2.4 Two-Stage Images Generation

Two-Stage Image Generation is a method that decomposes the image creation process into two distinct phases. This approach offers enhanced control and

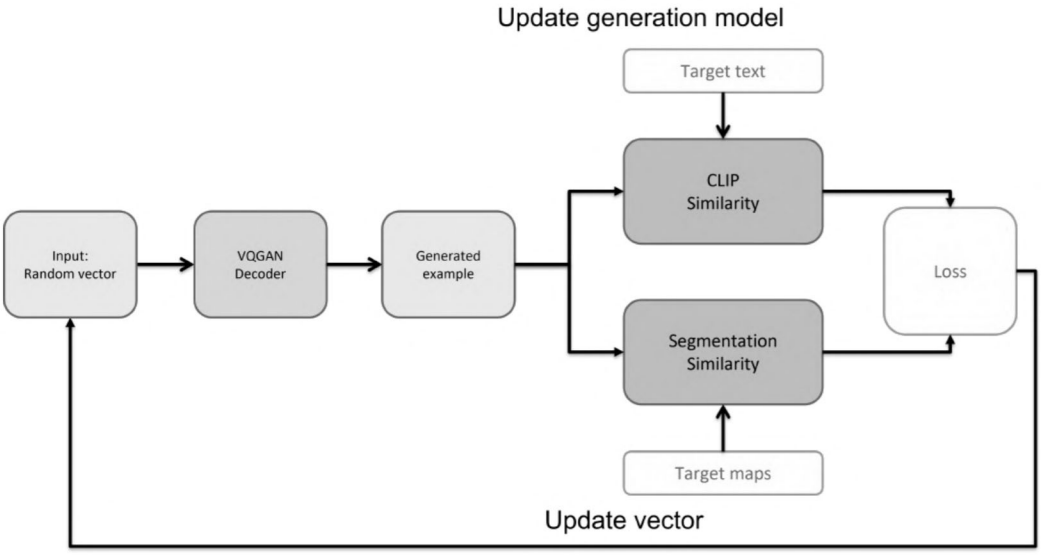


Figure 2.3: The controllable image generation process of a TCIG architecture

refinement, leading to higher quality and more accurate results compared to single-stage methods. In the two-stage image generation process, the first stage typically involves generating a rough sketch or outline of the desired image based on an initial input, such as a text prompt. This intermediate output provides a foundational structure that captures the key elements and layout of the final image. In the second stage, the initial sketch is refined and detailed, incorporating additional inputs and adjustments to produce the final high-quality image.

Zhang et al. introduces a novel two-stage framework for automatic sketch colorization [49]. The first stage involves generating an initial colorization of the sketch using a coarse-to-fine approach, where the initial colors are predicted based on the sketch. The second stage refines these colors to produce high-quality, natural results. The method effectively divides the complex task of colorization into two simpler tasks, improving the learning process and the final colorization quality. This work demonstrated the effectiveness of this approach through various experiments and comparisons with existing methods, showing superior performance in terms of color accuracy and visual appeal. The initial stage relies heavily on predicting colors from the sketch, which can sometimes be inaccurate, especially for sketches with ambiguous or incomplete details. This can lead to initial colorization that require significant refinement in the second stage.

The TCIG framework also proposed a two-stage approach for controlled image generation [50]. The first stage generates an initial image based on

input sketches and text descriptions, guided by pre-trained segmentation models. As shown in Figure 2.3, the initial input consists of a random vector, which is passed through a VQGAN decoder to produce a generated example. During this process, the framework compares the generated image with target text and segmentation maps by calculating CLIP similarity and segmentation similarity, respectively. These comparisons serve as key components of the loss function, helping the model iteratively improve the image generation by aligning it with the textual and segmentation conditions. The second stage employs a pre-trained diffusion model to enhance and refine the image's quality.

This method effectively combines the strengths of segmentation guidance and diffusion models, enabling the generation of high-quality, controllable images without extensive fine-tuning. The approach is flexible, allowing for the use of various diffusion methods, and significantly improves the generated images' realism and detail. However, the quality of the final image heavily relies on the accuracy of the initial segmentation. Errors or inaccuracies in the segmentation stage can propagate through to the final image, affecting overall quality.

TexControl is a two-stage model specifically designed for fashion image generation from sketches [51]. In the first stage, the model uses ControlNet to generate outline previews from input sketches. The second stage refines these previews using an image-to-image model to add detailed textures and specified materials, completing the fashion design process. This method enables precise control over the generated textures and materials, making it highly suitable for fashion design applications. The paper shows that TexControl can generate more accurate and complex materials compared to other models, providing a significant improvement in the quality and usability of generated fashion images. However, while TexControl can generate detailed textures and materials, it may struggle with very complex or unconventional textures that were not well-represented in the training data. Besides, ensuring consistency in the generated textures and materials across different parts of the garment can be challenging, especially when dealing with complex designs that require precise and consistent detailing.

The studies discussed in this section explore various two-stage approaches for image generation, with a focus on controlling and refining outputs by breaking down the task into distinct phases. Zhang et al. introduces a method for sketch colorization, where the initial rough colorization is generated first, followed by a refinement stage to improve the quality. The TCIG framework follows a similar approach, combining segmentation guidance and diffusion models to refine images over two stages. TexControl applies a two-step process to generate detailed fashion images from sketches, with the

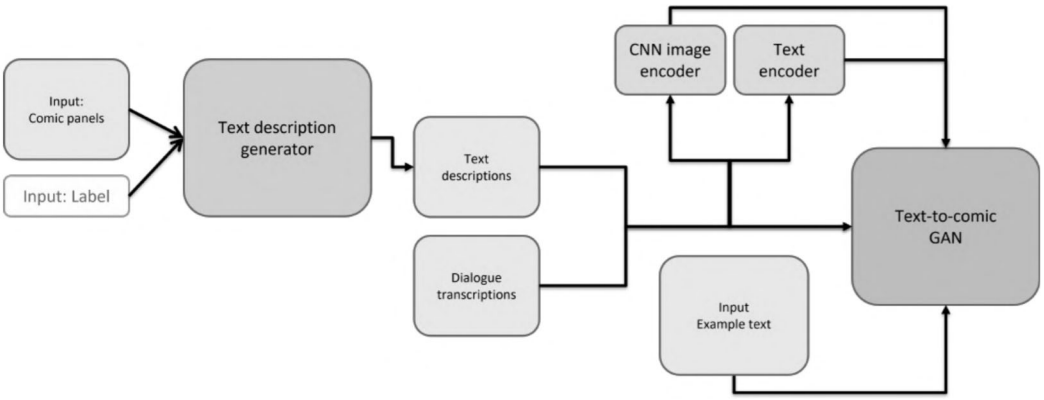


Figure 2.4: The pipeline of ComicGAN.

first stage generating an outline and the second stage refining textures and materials.

These two-stage frameworks relate directly to the theme of this paper, as our proposed model also adopts a two-stage approach. In our case, the first stage generates sketches based on text prompts, providing users with the ability to modify and arrange these sketches. In the second stage, the model refines the modified sketches to generate high-quality, coherent image sequences. By utilizing this two-step process, similar to the frameworks discussed, our model ensures higher control over the image generation process and produces more accurate results that align with user input, addressing the limitations of single-stage models.

2.5 Consistent Images Generation

Consistent images refer to a series of images that maintain consistency in terms of content, style, and features across multiple frames or instances. Consistent images are essential for maintaining viewer immersion and ensuring the clarity of the narrative or visual message. Inconsistency can lead to distracting artifacts that break the viewer's engagement and undermine the storytelling. Consistent image generation has significant applications in various fields, including animation [52] [53], video production [54] [55] and gaming [56] [57].

2.5.1 ComicGAN

ComicGAN presents an innovative approach to generating manga-style images from ordinary photographs [58]. The model utilizes a GAN-based archi-

tecture to capture the unique artistic style of manga and applies it to convert real-world images into manga-style illustrations. However, this method is highly dependent on high-quality images as input, which limits its flexibility in more imaginative or abstract contexts where real-life photographs are unavailable.

As shown in Figure 2.4, the ComicGAN framework includes multiple components: a text description generator that generates descriptions and transcriptions from comic panels and labels, followed by a Text-to-Comic GAN that uses both CNN and text encoders to process input text and images. The architecture allows the model to learn the textual and visual relationships needed to generate accurate comic-style images based on both input descriptions and example images. This two-stage process enables more detailed control over the generation process, aligning the final outputs with both visual and textual input data.

However, the model's reliance on high-quality real-world images and a large, curated dataset for training presents scalability challenges, especially when users aim to create manga illustrations from abstract or nonexistent scenes.

2.5.2 MangaGAN

MangaGAN introduces an innovative approach to generating manga-style images from photographs without requiring paired training data [59]. MangaGAN circumvents the requirement of a bunch of paired training data by employing a CycleGAN architecture, which allows the model to learn the mapping between photos and manga styles through unpaired image sets. The introduced method significantly reduces the data preparation effort and makes the model more versatile in various applications. Despite the advancements, MangaGAN focus on specific manga styles, which limits its generalizability. The model performs well with the manga style it was trained on but may not effectively adapt to other artistic styles without significant retraining.

In recent years, researchers apply diffusion model to consistent image generation task as well. Jeong et al. generates coherent storybooks from plain text input [60]. The method combines Large Language Models (LLMs) and text-conditioned Latent Diffusion Models (LDMs) to generate images without additional training data. Another contribution is the injection of iterative coherent identity. Face restoration models are applied to the initial images to enhance quality and address facial feature mutations. However, when the model faces complex or ambiguous text descriptions, it may struggle with intricate or abstract descriptions that require deeper contextual

understanding.

2.5.3 StoryLDM

Make-A-Story introduces a novel approach to generating coherent and consistent visual narratives by leveraging a visual memory module [61]. This method aims to produce a sequence of images that maintains consistency in characters, backgrounds, and overall story elements throughout the generated story. The key innovation is the integration of crucial visual information across the sequence of images through a visual memory module. This module ensures that visual elements, such as character appearances and background details, remain consistent throughout the narrative. By referencing previously generated frames, the model maintains coherence and continuity in the story. However, the complexity of the visual memory module can lead to increased computational requirements and longer processing times. Additionally, the model's reliance on accurate memory retrieval can pose challenges, especially in scenarios with highly dynamic or complex narratives, where maintaining consistency becomes more difficult.

2.5.4 The Chosen One

The Chosen One addresses the challenge of maintaining character consistency in text-to-image generation tasks [62]. Within the diffusion framework, The Chosen One introduces a character consistency module designed to track and preserve the distinct visual features of characters across multiple generated images. The process begins by generating a large set of images using a text-to-image model based on the input prompt. These images are embedded into a high-dimensional feature space, and the embeddings are clustered using the K-MEANS++ algorithm. Small clusters are filtered out, and the most cohesive cluster, characterized by the least variance among its members, is selected. This selection is followed by a personalization process that updates the model's weights and text embeddings to better capture the consistent identity of the character. By embedding character-specific attributes and enforcing consistency constraints during the diffusion process, the model ensures that characters remain recognizable and stable throughout the sequence. This approach effectively maintains character identity, enhancing the coherence and continuity of the generated visual narratives. One notable drawback of The Chosen One is its limitation to maintaining the consistency of only the main character. While the primary character's appearance remains stable, other elements of the scene, such as backgrounds and secondary characters, may still undergo significant changes.

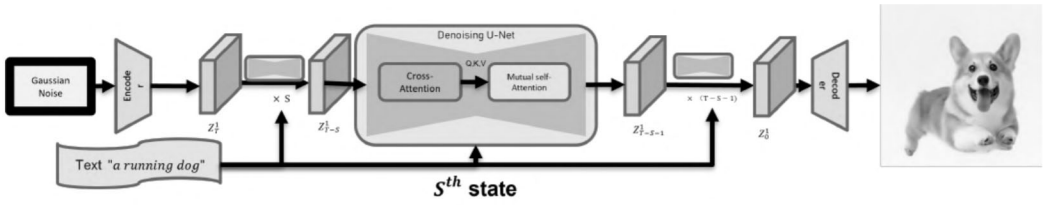


Figure 2.5: The network structure of MasaCtrl.

2.5.5 MasaCtrl

MasaCtrl introduces an innovative approach for enhancing consistency in image editing tasks by mutual self-attention within image generation models [63]. The method leverages a mutual self-attention mechanism, as shown in Figure 2.5. One of MasaCtrl's core contributions is its fine-grained control within mutual self-attention layers, ensuring that key image features remain consistent throughout the editing or generation process. This approach is especially valuable in sequential image generation applications, such as animation and comic storyboard creation, where preserving continuity across frames is essential.

First, Masactrl performs a diffusion inference on a real image, capturing and storing intermediate attention information from each layer as reference data. During the subsequent image editing phase, the spatial transformer layer's mutual self-attention mechanism is guided by this previously stored attention information, using it as a conditioning input. This setup allows the model to treat the original attention data as a "ground truth" condition, guiding the new image generation to adhere closely to the visual characteristics of the original image.

This tuning-free, dynamic approach makes MasaCtrl computationally efficient and suitable for real-time applications. It addresses a significant limitation in existing diffusion-based models, which often require substantial re-training or parameter adjustments to achieve similar levels of consistency across frames.

MasaCtrl, while offering valuable advancements in maintaining consistency across image edits, has several limitations. Firstly, the image quality produced by MasaCtrl tends to be relatively lower than other state-of-the-art methods. This quality constraint can limit its usability in applications that require high-fidelity visuals. Secondly, when applied to sequential image generation tasks, MasaCtrl's consistency mechanism primarily relies on the previous frame in the sequence. This approach has a significant drawback where any deviation or error in a single frame can propagate

through the subsequent frames, leading to a cascade of inconsistencies across the sequence.

Furthermore, MasaCtrl operates as a single-stage generation model. If the generated image does not fully align with the user's intentions, the user must either modify or regenerate the entire sequence to correct the discrepancy. This lack of intermediate control means users cannot make adjustments within the generation process, which can be time-consuming and inefficient, especially for lengthy sequences or complex visual narratives. Consequently, MasaCtrl may be less suited for applications where iterative refinement or user-directed adjustments are crucial for achieving the desired

To be notice that MasaCtrl's approach to maintaining consistency differs significantly from the method proposed in this research, although they might appear similar at first glance. Firstly, while MasaCtrl is primarily designed for real-image editing tasks, it can be adapted for image generation but does not inherently focus on it. In contrast, the model presented in this research is developed specifically for image generation tasks, where ensuring consistency throughout a sequence of generated images is essential.

The methods for maintaining image consistency also differ fundamentally. MasaCtrl depends on a mutual self-attention mechanism that leverages information from the preceding image, meaning each image only considers the immediately previous frame. This reliance can lead to issues when errors or inconsistencies appear in any single frame, as they will likely propagate through the entire sequence. However, the Multi Aggregation Attention (MAGA) mechanism introduced in this research differs by maintaining consistency across all preceding images in the sequence. This approach significantly reduces the risk of cascading errors, addressing one of MasaCtrl's limitations.

Finally, the proposed method is built on a two-stage framework, allowing users to make modifications to sketches as intermediate results. This offers high flexibility and control, enabling users to adjust or modify freely within the generation process itself. MasaCtrl, as a single-stage method, lacks this level of iterative refinement, meaning that users would need to regenerate the entire sequence if the output does not meet their expectations. This two-stage method not only enhances user control but also addresses key drawbacks observed in single-step models like MasaCtrl.

This section explores various research works focusing on consistent image generation, particularly for applications in storytelling, manga creation, and animation. Studies like MangaGAN and ComicGAN highlight the importance of consistency when generating manga-style illustrations, emphasizing the challenges of maintaining visual coherence when transforming real-world images into stylized artworks. Moreover, Make-A-Story and The Chosen One

delve into generating coherent visual narratives by integrating visual memory modules and character consistency mechanisms, ensuring that characters and environments remain visually stable across sequences of images.

The common thread across these works is the need for consistency, particularly when generating multiple frames or sequential images. This is crucial for maintaining the integrity of the story or message conveyed through visual media. In relation to our work, these studies underscore the importance of feature consistency when generating a sequence of images from text prompts or sketches. Our approach, which involves a two-stage generation process, aims to further improve this consistency by allowing users to modify and refine sketches during the generation process, ensuring better coherence in the final output. This is particularly relevant when generating comic or animation sequences, where consistency in character features, poses, and background elements is critical for storytelling and visual engagement.

Chapter 3

Proposed Method

This chapter provides a comprehensive discussion of the proposed image generation methodology. Firstly, in Section 3.1, we introduce the LDM pipeline, which enables efficient and high-quality image generation by operating within the latent space, thereby reducing computational complexity without compromising output fidelity. We also explore the core mechanisms of LDM, such as the forward diffusion process and reverse denoising, which lay the groundwork for understanding the enhancements introduced in our method. Section 3.2 presents an overview of the overall process of the proposed method. Section 3.3 details the adjustment and modification of intermediate results. Section 3.4 explains the sketch generation process during the first stage. Finally, Section 3.5 details the proposed architecture of the second stage.

3.1 Latent Diffusion Model

Our proposed model utilizes the Latent Diffusion Model (LDM) as its backbone [17], making it essential to provide a detailed overview of LDM in this chapter. LDM has proven to be a powerful framework for various image generation tasks, offering robustness and flexibility critical for the development of advanced generative models. In Section 3.1.1, we introduce the pipeline of LDM. In Section 3.1.2, we discuss the attention mechanism within LDM. In Section 3.1.3, we explore the U-Net backbone.

3.1.1 Network Structure

The Latent Diffusion Model (LDM) [17] operates by mapping input data into a latent space where the diffusion process is performed. This approach significantly reduces computational complexity compared to working directly in the high-dimensional image space. The image generation process in LDM can be broken down into several key steps: encoding the image into a

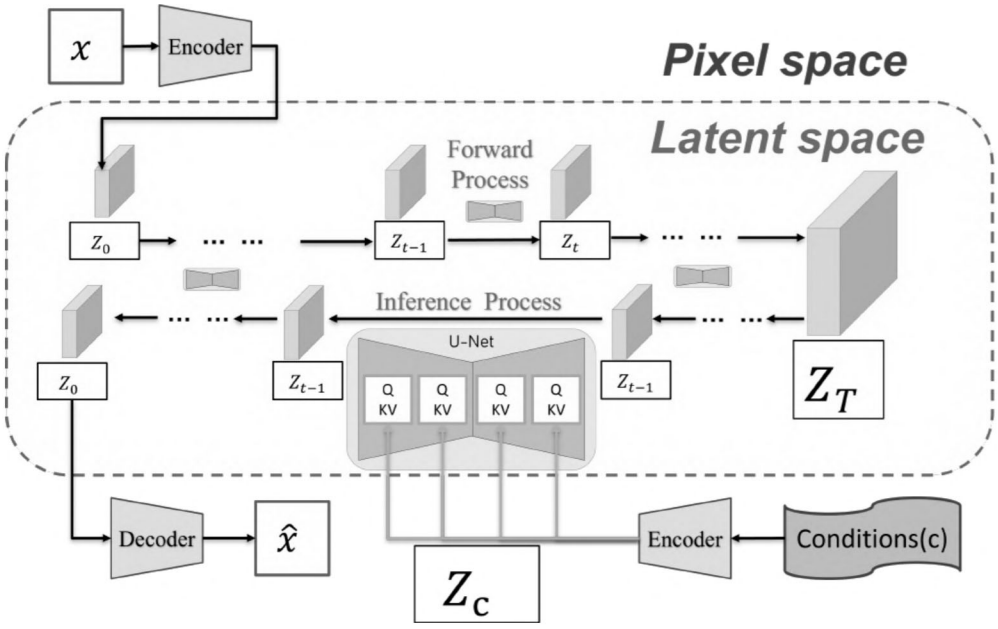


Figure 3.1: The structure of latent diffusion model.

latent representation, applying the diffusion process, and decoding the latent representation back into an image. In Figure 3.1, each step is detailed.

During a training process, the first step involves encoding the input image x into a latent space representation Z_0. This is typically achieved using an encoder network:

$$Z_0 = E(x) \tag{3.1}$$

The encoder network compresses the high-dimensional image x into a lower-dimensional latent vector Z_0, capturing essential features while reducing computational load.

At the same time, conditioning plays a crucial role in the LDM framework, allowing the model to incorporate additional information to guide the generation process. Conditions can include various forms of auxiliary data, such as text prompts, sketches, or other contextual information. This conditioning is typically integrated into the model through concatenation or attention mechanisms.

For instance, if a textual description c is used as a condition, it can be encoded into a latent vector Z_c using a text encoder:

$$Z_c = E(c) \tag{3.2}$$

Figure 3.2: Photo-realistic images generated by Imagen [1].

Once the image is encoded into the latent space, the diffusion process begins. The diffusion model iteratively adds and removes noise to refine the latent representation. This process can be described using a series of latent variables Z_t for $t = 0, 1, \ldots, T$, where T is the total number of diffusion steps.

The forward diffusion process, which adds noise ϵ_t, can be defined as:

$$\mathbf{Z}_t = \sqrt{\alpha_t}\mathbf{Z}_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{1}) \tag{3.3}$$

where α_t is a variance schedule controlling the amount of noise added at each step t.

The inference process, which denoises the latent representation, is parameterized by a U-Net D_0 and is defined as:

$$\mathbf{Z}_{t-1} = D_\theta\left(\mathbf{Z}_t, t, \mathbf{Z}_c\right) \tag{3.4}$$

During training, the model learns to predict the noise ϵ_t added in the forward process, enabling it to denoise the latent representation effectively:

$$\epsilon_\theta\left(\mathbf{Z}_t, t, \mathbf{Z}_c\right) = \frac{\mathbf{Z}_t - \sqrt{\alpha_t}\mathbf{Z}_{t-1}}{\sqrt{1-\alpha_t}} \tag{3.5}$$

The training objective is to minimize the following loss function, which measures the difference between the predicted and actual noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{Z}_t, t, \mathbf{Z}_c}\left[\left\|\epsilon_t - \epsilon_\theta\left(\mathbf{Z}_t, t, \mathbf{Z}_c\right)\right\|^2\right] \tag{3.6}$$

After the latent representation has been refined through the diffusion process, it is decoded back into the image space using a decoder network D:

$$\hat{\mathbf{x}} = D\left(\mathbf{Z}_0, \mathbf{Z}_c\right) \tag{3.7}$$

The decoder network transforms the latent vector Z_0 back into an image $\hat{\mathbf{x}}$, ideally resembling the original input image x but with the desired generative modifications, influenced by the conditional information.

After the training is done, the model is able to inference image. The first step involves preparing the conditional vector Z_c. For instance, a text description c is encoded into Z_c using the text encoder E_c as Formula 3.2.

At the same time, A latent vector Z_T is sampled from a Gaussian distribution:

$$Z_T \sim \mathcal{N}(0, \mathbf{1}) \tag{3.8}$$

Then, the latent vector Z_t is iteratively refined as Formula 3.5 At each step t, the model predicts the noise and refines the latent vector using the conditional information.

At last, the refined latent vector Z_0 is decoded into the image space using the decoder network D under the instruction of Formula 3.7.

By utilizing a trained LDM, high-quality images can be generated with or without conditions. Examples of the results are shown in Figure 3.2

3.1.2 Attention

Attention mechanisms play a crucial role in LDM by allowing the model to selectively focus on different parts of the input data [64]. This selective focus enhances the model's ability to capture relevant features and dependencies, thereby improving the quality and coherence of the generated images. In LDM, attention mechanisms are particularly effective when combined with conditional inputs, as they enable precise control during generation process.

Attention mechanisms play a crucial role in LDM by allowing the model to selectively focus on different parts of the input data. This selective focus enhances the model's ability to capture relevant features and dependencies, thereby improving the quality and coherence of the generated images. In LDM, the integration of attention mechanisms within U-Net helps to ensure that the conditional information is effectively utilized throughout the network.

Attention mechanism can be categorized into self-attention and cross-attention. Self-Attention allows the model to weigh the importance of different parts of the same latent representation. While Cross-Attention allows the model to gain the information of the condition relative to the latent representation.

Let $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively. These matrices are derived from the input features and the conditional input. The attention mechanism computes the attention weights and applies them to the values to produce the output:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.9}$$

where d_k is the dimension of the key vectors, used to scale the dot product for numerical stability.

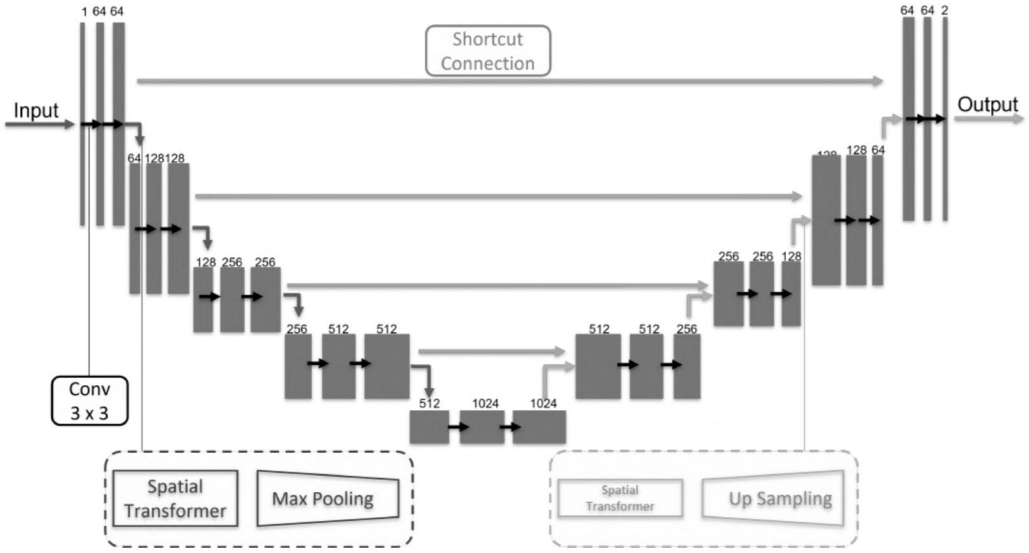


Figure 3.3: The architecture of U-Net in Latent Diffusion Model

To incorporate conditional information, cross-attention mechanism is modified to handle both the latent representation and the conditional input. Suppose $\mathbf{Z}$ is the latent representation and $\mathbf{Z}_c$ is the conditional vector. The cross-attention mechanism integrates these inputs as follows:

$$Q = W_Q^{(i)} \cdot z_t, K = W_K^{(i)} \cdot z_c, V = W_V^{(i)} \cdot z_c \tag{3.10}$$

Notably, in i^{th} layer, $W_Q^{(i)}$, $W_Q^{(i)}$ and $W_Q^{(i)}$ represent learnable matrices.

3.1.3 U-Net

The LDM uses the U-Net architecture as its backbone due to its effectiveness in capturing both local and global features through its encoder-decoder structure with skip connections [65]. U-Net is a convolutional neural network (CNN), designed for image segmentation but has been adapted for various image generation tasks due to its powerful feature extraction capabilities.

U-Net is an Encoder-Decoder network, which consist of two main parts. Encoder captures the context of the input image by progressively reducing its spatial dimensions while increasing the number of feature channels, while Decoder reconstructs the image by progressively increasing its spatial dimensions while combining low-level features from the encoder through skip connections.

As illustrated in Figure 3.3, the encoder down-samples the input image and extracts high-level features. Convolutional Layers are used for applying convolution operations to extract features from the input image. Max Pooling layers reduce the spatial dimensions by taking the maximum value over a sliding window, effectively down-sampling the input. The decoder is designed for up-sampling the features back to the original image resolution. Convolutional Layers process the concatenated features to refine the image, while Transposed Convolutional Layers perform up-sampling, increasing the spatial dimensions of the feature maps.

Shortcut connections play a crucial role in U-Net by connecting corresponding layers. Shortcut connections allow the model to combine low-level and high-level features, ensuring that the final output retains fine details while incorporating contextual information, which is essential for generating high-quality images that maintain both local and global consistency

Spatial Transformer layers (Cross-Attention layers) are typically inserted between the convolutional layers within both the encoder and decoder. Cross-attention layer first takes the latent representation from the previous convolutional layer and the conditional vector. The query matrix is formed from the latent representation, while the key and value matrices are derived from the conditional vector. Next, the attention weights, which determine the importance of different parts of the conditional information is computed. These weights are then used to integrate the conditional information into the latent representation, ensuring that the generated images adhere to the specified conditions while maintaining high quality and coherence.

3.2 Two-stage Image Generation

Current One-step text-to-image generation models do not allow users to modify images during the generation process [58] [59] [44] [45] [31] [30] [43] [27] [32], resulting in final images that often fail to meet the users' expectations. Besides, as noted in Chapter 2.3, current sketch-based image generation models require high-quality sketches with clear and precise lines to produce accurate and coherent images [47] [48] [44] [45] [46] [66]. The quality and clarity of the input sketches may directly affect the generated output. Ambiguous, incomplete, or poorly drawn sketches can lead to sub-optimal results. Additionally, users often face issues like a lack of flexibility in input modification, inconsistencies in generated details, and even computational challenges, which complicate the generation process. Furthermore, common problems like mode collapse or difficulties in handling intricate designs can arise, leading to reduced diversity in outputs and unintended deviations from

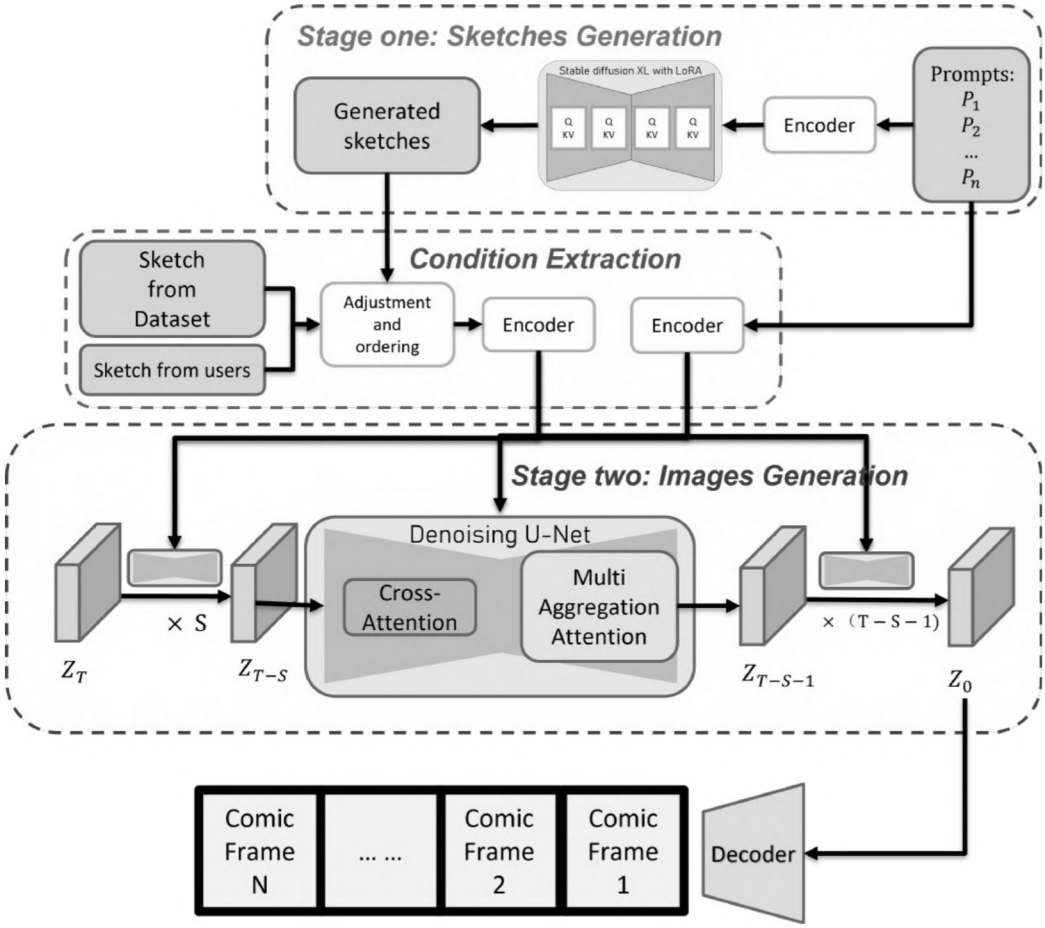


Figure 3.4: The framework of proposed model.

the intended structure. These limitations collectively hinder the accessibility and practicality of current models, particularly for non-expert users or those working with complex or imperfect inputs.

To address these limitations, we propose a method that introduces a flexible, two-stage generation process, enabling users to make modifications during the image generation process. By providing these options, we lower the barrier to entry and make the technology accessible to a wider audience. The proposed method involves several key Stage, as illustrated in Figure 3.4.

In the Sketch Generation Stage (Stage 1), users begin by inputting several text prompts to initiate the process. These text prompts are encoded using the Contrastive Language-Image Pre-Training (CLIP) model [67]. Users provide descriptive text prompts that detail the elements and characteristics. After that, text prompts are encoded using the CLIP model's encoder. CLIP

is designed to understand and map textual descriptions into a latent space that is compatible with image features. This encoding process generates textual feature representations that align with the sketch features encoded by Canny and VAE. Based on the encoded textual descriptions, the model generates corresponding sketches, providing an initial visual representation guided by the user's input.

3.3 Condition Extraction

In the Condition Extraction stage, text prompts provided by the user are encoded to guide the image generation process. As depicted in Figure 3.4, after the user inputs text descriptions, these text prompts are processed using the CLIP model (Contrastive Language-Image Pretraining). The CLIP model is specifically designed to understand and map textual descriptions into a latent space that is compatible with image features. In this way, the encoded text prompts serve as a condition to inform the sketch generation and refinement processes.

3.3.1 Condition Selection and Integration

Once the text prompts are encoded, they are combined with other conditions like sketches from Stage 1, pre-existing sketches from the dataset, or user-provided hand-drawn sketches, to create a comprehensive condition for the final image generation in Stage 2. The encoded text conditions are input into the latent diffusion model, which utilizes these features during the iterative denoising process. In this process, the text encoding interacts with the visual features through a cross-attention mechanism in the U-Net architecture, ensuring that the final images are aligned with the user's textual descriptions and maintain consistency throughout the sequence. This approach allows for a flexible and powerful mechanism to generate images that closely match user intent, as informed by both textual and visual inputs.

Figure 3.5 illustrated the process where users choose sketches to proceed. Users have three options for selecting sketches to form the foundation of their final image sequence. Firstly, users may use the sketches generated in Stage 1, based on text prompts they provided. If they are unsatisfied with these sketches, they can modify the text input and regenerate sketches until the desired output is achieved. Secondly, the system offers a curated sketch dataset with pre-made still-life and character sketches, which users can select to include in the sequence. Lastly, for users with artistic skills, there is an option to integrate their own hand-drawn sketches directly into the sequence.

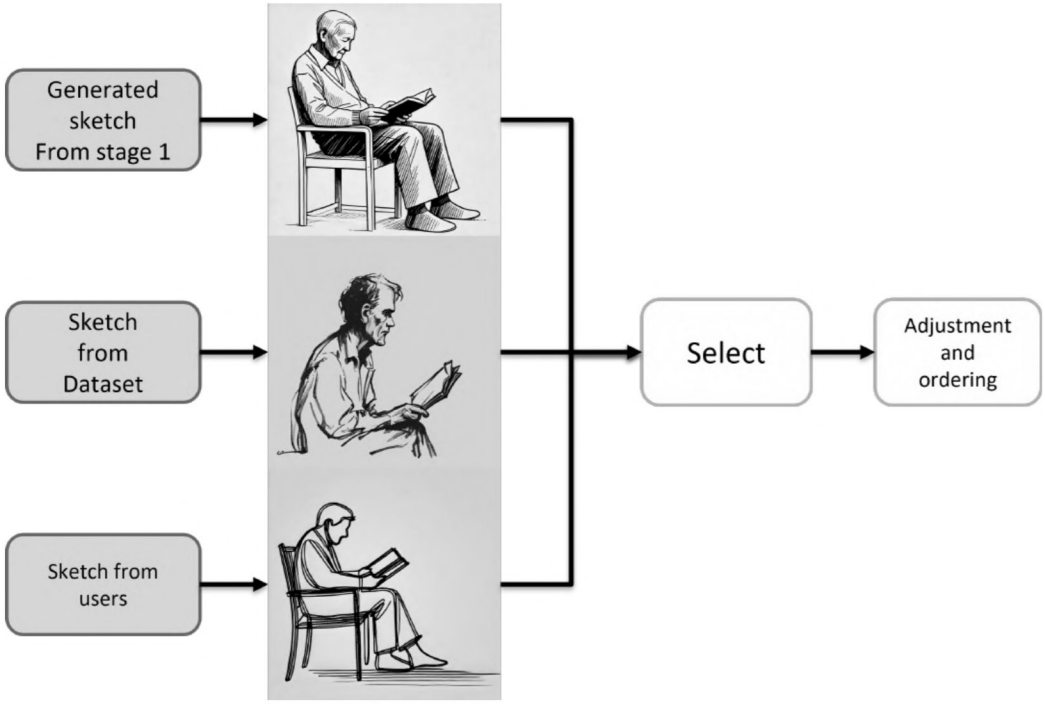


Figure 3.5: Sketch selection process in condition extraction stage, where users can choose between sketches generated in Stage 1, pre-existing sketches from a dataset, or their own hand-drawn sketches.

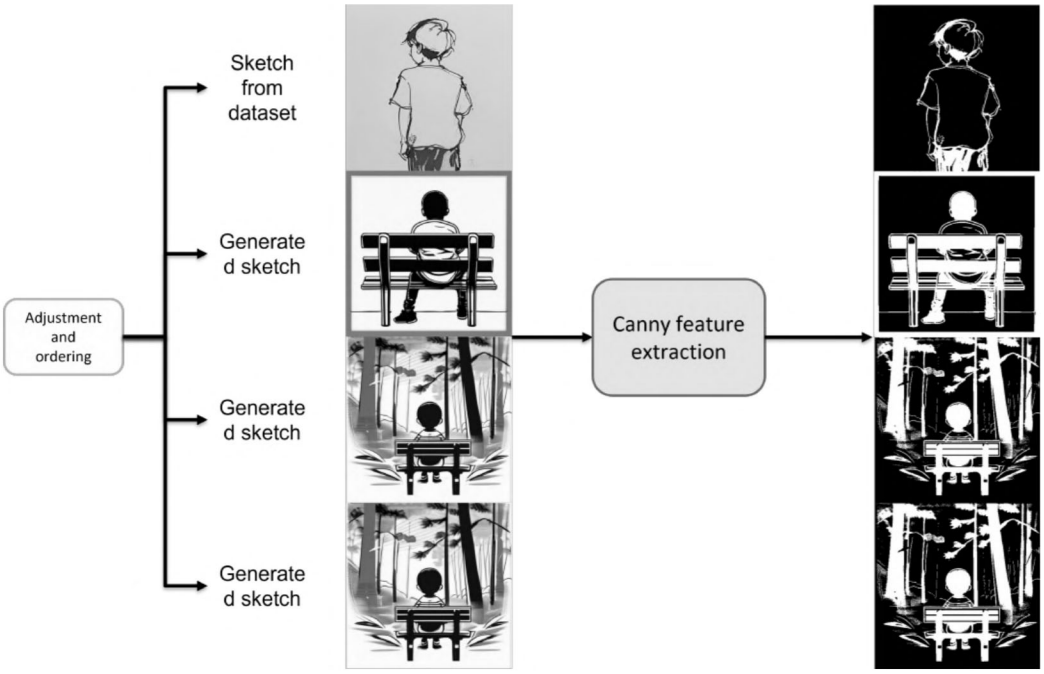


Figure 3.6: Canny edge detection, used to extract features from the sketches, are then encoded to guide the image generation process.

Once selected, users can further adjust and arrange the sketches in the desired order to ensure consistency and coherence in the final generated sequence in Stage 2.

3.3.2 Condition Extraction by Canny

Once the user modifications and ordering of the sketches are completed, the Canny edge detection model, inspired by ControlNet [26], is employed to extract key features from the sketches [68], as shown in Figure 3.6. Canny extracts the prominent edges and outlines of the sketches, capturing the essential structural elements. These extracted edge features are then encoded using a Variational Autoencoder (VAE) [69], which captures the latent representations of the sketch features. This ensures that the representations effectively guide the subsequent stages of image generation.

The sketch features and text features encoded in Stage 1 are used as conditions input into the proposed latent diffusion-based model, Stage 2, ensuring that the output images adhere closely to the user's specifications. The model starts with an initial noisy image Z_T and iteratively refines it. At each step, the previously generated image is reintroduced into the process, along

with new sketch features and text inputs. The cross-attention mechanism within the U-Net uses the encoded sketch and text features to guide the denoising process, ensuring that the generated layout maintains coherence and accurately represents the input conditions. The Multi Aggregation Attention mechanism within the U-Net ensures consistency in character features and other elements across the sequence of images. This iterative approach, denoted by the transition to Z_0, produces a sequence of coherent and high-quality images that reflects the user's intent.

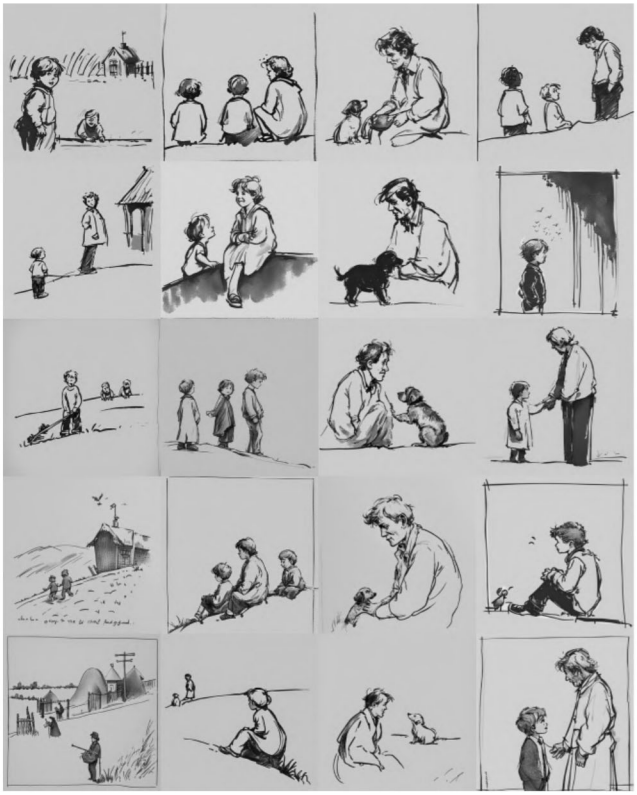
3.3.3 Dataset: SketchXL

Current sketch datasets [46] share a common limitation: they consist of individual sketches without consistency across a series of sketches. This lack of sequential consistency limits their applicability for tasks that require a coherent series of sketches. Therefore, this research recognizes the necessity to create a new dataset, SketchXL, which will support the development of models that can generate and handle sequences of sketches. By allowing users to select and order sketches from these consistent sketch groups, SketchXL will directly support the development and implementation of the advanced image generation workflow outlined in this paper.

The SketchXL dataset consists of two main parts: the Story Group and the Sundry Group, both of which contain sketches generated using Stable Diffusion XL to ensure consistent quality and style across all samples. This is important for enabling both users and models to handle a coherent sequence of sketches in their workflows.

The Story Group contains a collection of 50 distinct stories sketches and their text prompts. Each story is divided into four sequential parts, and each part contains 80 sketches, resulting in a total of 16,000 sketches. This group emphasizes narrative and thematic consistency, ensuring that the transitions between frames are smooth and logical, as shown in Figure 3.7. These sketches are particularly useful for applications that require coherent storytelling, where maintaining consistency in characters, backgrounds, and other elements across multiple frames is vital.

In contrast, the Sundry Group provides a set of 1,000 sketches of commonly used objects, as depicted in Figure 3.8. Unlike the Story Group, the Sundry Group does not focus on maintaining consistency across sketches. Instead, it offers a variety of standalone sketches, representing diverse everyday items, making this group useful for users who may need individual objects to complement or enhance their sequences. The sketches in the Sundry Group can also serve as placeholders or additional visual elements in broader narratives or complex visual compositions.



Children were at play in their play-ground one
day, when a herald rode through the town, blowing a trumpet, and crying aloud, "The King! the King passes by this road to-day. Make ready for the King!"
The children stopped their play, and looked at one another.
"Did you hear that?" they said. "The King is coming. He may look over the wall and see our playground; who knows? We must put it in order."
The playground was sadly dirty, and in the corners were scraps of paper and broken toys, for these were careless children. But now, one brought a hoe, and another a rake, and a third ran to fetch the wheelbarrow from behind the garden gate. They labored hard, till at length all was clean and tidy.
"Now it is clean!" they said; "but we must make it pretty, too, for kings are used to fine things; maybe he would not notice mere cleanness, for he may have it all the time."
Then one brought sweet rushes and strewed them on the ground; and others made garlands of oak leaves and pine tassels and hung them on the walls; and the littlest one pulled marigold buds and threw them all about the playground, "to look like gold," he said.
When all was done the playground was so beautiful that the children stood and looked at it, and clapped their hands with pleasure.
"Let us keep it always like this!" said the littlest one; and the others cried, "Yes! yes! that is what we will do."
They waited all day for the coming of the King, but he never came; only, towards sunset, a man with travel-worn clothes, and a kind, tired face passed along the road, and stopped to look over the wall.
"What a pleasant place!" said the man. "May I come in and rest, dear children?"
The children brought him in gladly, and set him on the seat that they had made out of an old cask. They had covered it with the old red cloak to make it look like a throne, and it made a very good one.
"It is our playground!" they said. "We made it pretty for the King, but he did not come, and now we mean to keep it so for ourselves."
"That is good!" said the man.
"Because we think pretty and clean is nicer than ugly and dirty!" said another.
"That is better!" said the man.
"And for tired people to rest in!" said the littlest one.
"That is best of all!" said the man.
He sat and rested, and looked at the children with such kind eyes that they came about him, and told him all they knew; about the five puppies in the barn, and the thrush's nest with four blue eggs, and the shore where the gold shells grew; and the man nodded and understood all about it.
By and by he asked for a cup of water, and they brought it to him in the best cup, with the gold sprigs on it: then he thanked the children, and rose and went on his way; but before he went he laid his hand on their heads for a moment, and the touch went warm to their hearts.
The children stood by the wall and watched the man as he went slowly along. The sun was setting, and the light fell in long slanting rays across the road.
"He looks so tired!" said one of the children.
"But he was so kind!" said another.
"See!" said the littlest one. "How the sun shines on his hair! it looks like a crown of gold."

Figure 3.7: Story group in SketchXL dataset. One single story is separated into 4 scenarios, each containing 80 sketches for users to choose.

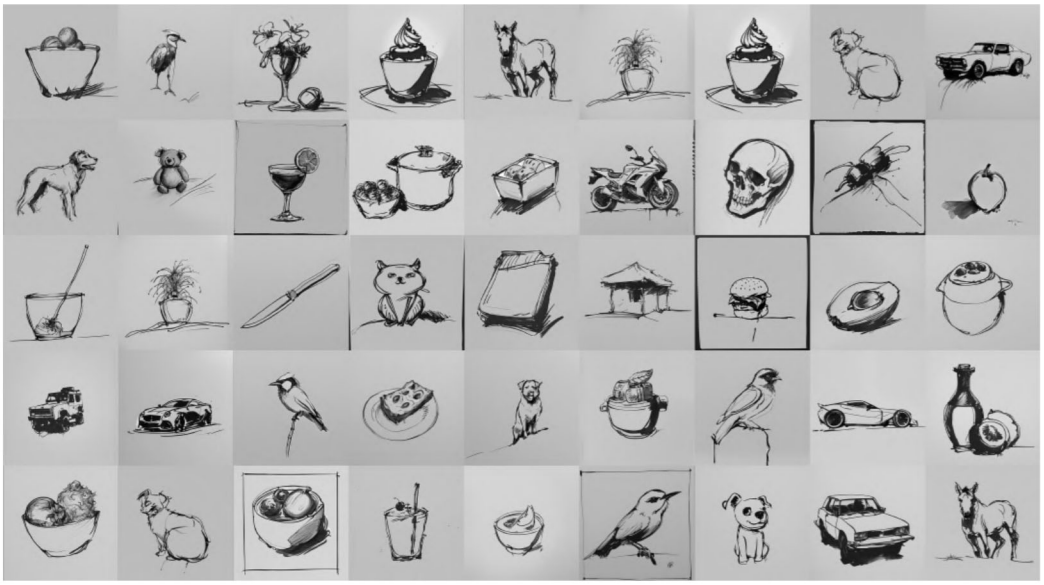


Figure 3.8: Sundry group in SketchXL dataset. 1,600 sketches are available for users to choose.

The high-quality sketches in SketchXL, generated using Stable Diffusion XL, are characterized by fine detail and clear lines, making them suitable for use as both rough drafts and final renderings. The structured nature of the Story Group, combined with the flexibility of the Sundry Group, makes the dataset highly versatile. This dataset provides a foundation that enables users to set out their sketches with precision and confidence, directly supporting the advanced two-stage image generation process proposed in this research. By allowing users to select and order sketches from coherent story-driven groups or diverse individual sketches, SketchXL enhances the accessibility and usability of image generation systems, empowering both novice and professional users to create detailed and consistent visual outputs efficiently.

3.4 Sketches Generation

In Stage 1 of our two-stage image generation process, we leverage the advancements of Stable Diffusion XL (SDXL) to enhance the efficiency and quality of sketch generation based on textual descriptions [70]. SDXL significantly enhances our ability to generate high-quality sketches from text prompts.

As Table 3.1 illustrated, SDXL offers significant improvements and ad-

Model	SDXL	Stable Diffusion v1.0
z-shape	$2048 \times 2048 \times 4$	$768 \times 768 \times 4$
Numbers of UNet Parameters	2.6B	860M
Channels	128	64

Table 3.1: Hyperparameters in Stable Diffusion XL compare with Stable Diffusion v1.0. z-shape represents the dimension of latent space.

vantages, compare to Stable Diffusion v1.0, particularly in the realm of image generation tasks. Firstly, SDXL features a larger parameter scale, enabling it to handle more complex and high-precision image generation tasks. Its enhanced architecture, incorporating more neurons and layers, results in superior image detail and overall consistency. This approach not only enhances the quality of the images but also increases the model's applicability across various scenarios. Despite requiring higher computational resources, SDXL's ability to produce detailed and high-quality images makes it an ideal choice for professional applications. In summary, SDXL demonstrates significant advancements in the field of image generation.

Our choice of SDXL for Stage 1, the text-to-sketch generation process, is driven by its unparalleled ability to produce detailed and high-quality sketches from textual descriptions. SDXL's increased parameter scale allows it to capture intricate details, which are crucial for generating accurate and coherent sketches. This capability ensures that the sketches generated in Stage 1 serve as a robust foundation for subsequent image refinement. Furthermore, SDXL's iterative denoising process in the latent space preserves the clarity and structural integrity of the sketches, facilitating seamless transitions into the next stages of image generation. By leveraging SDXL, we optimize the initial sketch generation process, ensuring high-quality outputs that enhance the overall effectiveness of our two-stage image generation method.

3.5 Image Generation

In the final image generation phase, the process begins with an initial noisy latent representation Z_T , which is progressively refined through a series of denoising steps using a U-Net architecture. The core of this process relies on the previously extracted conditions from the Condition Extraction phase, where both text and sketch features have been encoded. These encoded conditions are continually reintroduced throughout the denoising steps to

guide the model toward producing coherent and high-quality images.

As the denoising process unfolds, the latent representation is transformed iteratively. At each step, the model uses a cross-attention mechanism to focus on the relevant parts of the input sketches and text conditions, ensuring that the generated images reflect the user's input accurately. This attention mechanism helps the model to maintain the structural integrity and style of the sketches while incorporating the specific details provided in the text prompt.

In addition to the cross-attention, the model employs the Multi-Aggregation Attention (MAGA) Module, which plays a critical role in maintaining consistency across the image sequence. MAGA introduces an external memory component that aggregates information from the current and previous steps, comparing them with the input conditions. This ensures that key visual elements, such as character features and poses, remain consistent throughout the sequence of images, addressing potential issues of feature drift or inconsistency.

The denoising process continues iteratively, reducing the noise from Z_T to Z_0, the final latent representation. Once this process is complete, the model's decoder converts the latent variables back into pixel-space, generating the final sequence of images. This approach, which integrates both sketch and text conditions throughout the denoising process, results in a series of images that not only adhere to the user's initial design intentions but also maintain visual consistency and coherence across multiple frames.

Chapter 4

Multi Aggregation Attention Module

In this chapter, we introduce the Multi Aggregation Attention (MAGA) module, which plays a crucial role in addressing one of the fundamental challenges in image generation, particularly for sequential images, such as those used in comics, animation, and other visual storytelling tasks. Section 4.1 discussed the motivation for introducing MAGA into the proposed model, particularly its ability to ensure coherence across multiple frames by referencing information from earlier stages of the generation process. Next, in Section 4.2 explored the background and theoretical foundations of the MAGA mechanism, highlighting how it builds upon existing attention mechanisms and extends them by introducing memory aggregation. The technical implementation of MAGA is discussed in Section , covering its selective application within the denoising process and its role in maintaining image quality. Finally, Section concludes with a discussion of the optimal hyperparameter selection for MAGA and its potential for future research and experimentation.

4.1 Introduction

Maintaining temporal consistency is vital for creating coherent sequences where the key elements like character features, backgrounds, and other visual details remain consistent across multiple frames. Without such consistency, the viewer's experience can be disrupted, leading to a lack of immersion and narrative flow. While existing image generation models have made strides in generating high-quality images, many still struggle with maintaining coherence between sequential frames, often leading to visual inconsistencies. In the context of this work, the MAGA mechanism is essential due to the high demand for consistency across multiple frames in sequential image generation tasks. Traditional cross-attention mechanisms [64] used in latent diffusion models focus solely on the current input, making it difficult to maintain the

necessary coherence when generating a sequence of images.

For example, in visual storytelling, such as comic generation, the position of characters, their features, and the environment in which they are placed must remain consistent across a series of panels. Failing to maintain this consistency could lead to distracting discrepancies that affect the visual flow. While a standard cross-attention mechanism can ensure a certain level of feature fidelity for individual images, it does not account for the relationships between consecutive images in a sequence. This is where the MAGA module comes in, offering a method to reference previous frames, combining historical information with current inputs, and ensuring that the generated sequence remains coherent.

4.2 Background

The Multi Aggregation Attention (MAGA) mechanism is an attention-based technique designed to enhance the temporal consistency of image sequences by aggregating information across multiple stages of the image generation process. In a standard cross-attention block, the Query (Q), Key (K), and Value (V) are computed based on the current input and the corresponding conditions, such as a text prompt or sketch, to guide the generation of the image at each layer. However, this approach only considers the information within the current frame, limiting its ability to maintain coherence across multiple frames.

The MAGA mechanism extends this by introducing an external memory component that stores the QKV from previous stages of the image generation process. This memory allows the model to access and reuse relevant historical information when generating new frames, ensuring that the generated images align with both the current input and the information from earlier frames. This results in a much more coherent and temporally consistent output, which is crucial for tasks such as animation, where changes in character appearance or background inconsistencies can severely disrupt the viewing experience.

MAGA is inspired by similar concepts in story-driven diffusion models, such as storyLDM, which also addresses the need for consistency in narrative-driven image sequences [61]. However, MAGA takes this concept further by introducing selective memory aggregation, ensuring that it is applied only at critical stages of the generation process, as discussed in the technical sections that follow.

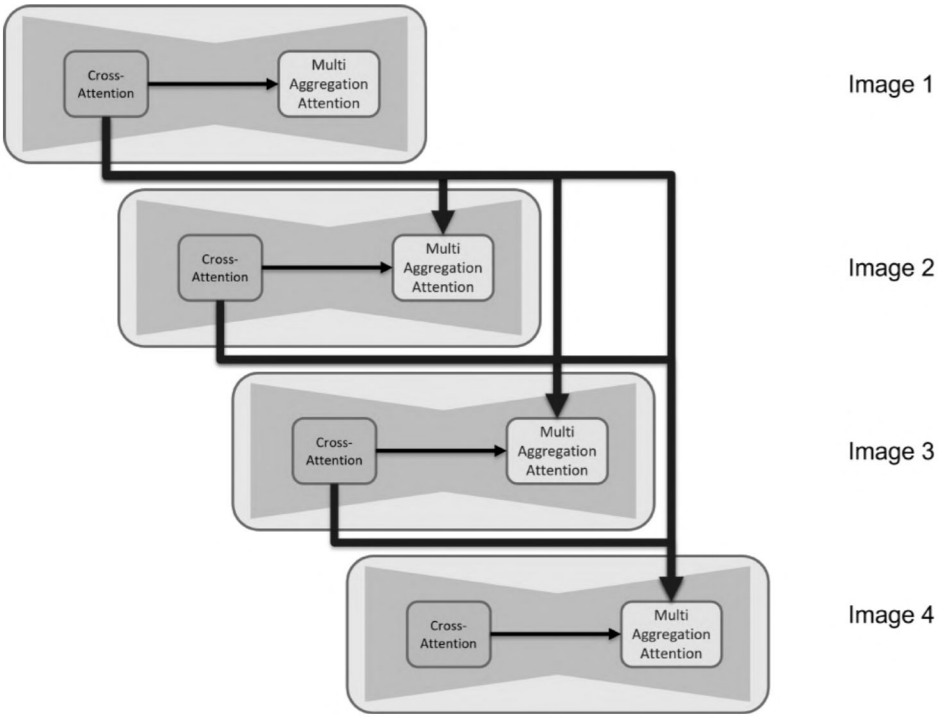


Figure 4.1: Information transfer between MAGAs during each generation process when generating multiple consistent images.

4.3 Mechanism

Cross-attention blocks in latent diffusion models typically process information from the current input. In each layer, the spatial feature output by the previous layer is used as the Query (Q), while the conditions at the same level are input as Key (K) and Value (V) as Formula 3.10. This mechanism ensures that the model attends to the relevant parts of the input conditions for each layer, thereby facilitating accurate and coherent image generation.

The Multi Aggregation Attention module (MAGA) introduces an external memory component that stores Query, Key, and Value (QKV) from the corresponding image generation process at the same position in previous levels of the network as illustrated in Figure 4.1. This allows MAGA to focus on the relationship between the current input and the information stored in memory, enabling the model to reason by combining historical context.

Specifically in Figure 4.2, cross-attention receives spatial feature from the previous layer is used as the Query (Q). The conditions corresponding to the current image n serve as both the Key (K) and Value (V). The cross-attention mechanism uses the Query to attend to the relevant parts of the

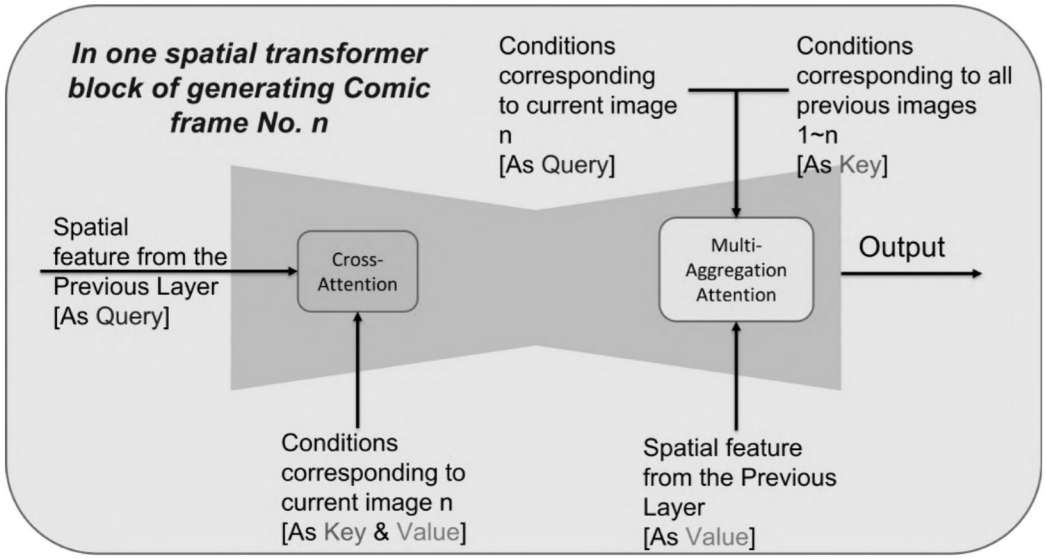


Figure 4.2: Information transfer inside one Spatial Transformer specifically during the generation process of one single image.

conditions (Key and Value) for the current image. In MAGA, The conditions corresponding to the current image n are again used as the Query (Q) as formula:

$$Q_{MAGA} = W_Q \cdot f(S^n) \tag{4.1}$$

All conditions corresponding to previous images are used as the Key (K) as formula:

$$K_{MAGA} = W_K \cdot f(S^{<n}) \tag{4.2}$$

The spatial feature from the previous layer is used as the Value (V) as formula:

$$V_{MAGA} = W_V \cdot f(Z^n) \tag{4.3}$$

MAGA enhances temporal consistency by storing Keys (K) from previous generation process.

However, as shown in Figure 3.4, in the early stages of the denoising process, the target image layout has not yet been fully formed. Introducing memory-attention control at these premature steps could disrupt the formation of the spatial layout of the target image. Early implementation of MAM can cause the model to overly rely on the historical context before the

overall structure of the image is established, resulting in an output that may not align with the intended design. In premature steps, the target image layout has not yet been formed, and performing MAGA control can disrupt the layout formation of the target image.

Thus, MAGAs are not implemented throughout the entire denoising process but is specifically introduced from the s^{th} denoising step to the final step. Additionally, within each denoising step, MAGAs are implemented only from the i^{th} spatial transformer block to the last one. This selective implementation strategy is crucial for maintaining the integrity and quality of the generated images.

Determining the optimal value of the hyperparameter s and i, which indicates the starting point for implementing MAGA, remains a topic for future research. Related ablation experiments will be illustrated and discussed in Section .

4.4 Ablation study

We introduced two critical hyperparameters for the network described in this paper: the specific layer at which the Multi-Aggregation Attention (MAGA) mechanism is incorporated and the number of inference steps. These hyperparameters are crucial for determining the optimal timing for introducing the MAGA mechanism into the inference process. Specifically, introducing MAGA too early can disrupt the formation of the spatial layout of the target image, leading to an overly reliance on historical context before the overall structure is established. This can result in outputs that do not align with the intended design. On the other hand, introducing the MAGA mechanism too late can result in insufficient information transfer between images, leading to a loss of consistency. To ascertain the optimal values for these parameters, we conducted a series of ablation studies. The results of these studies are presented in Figure 4.3, revealing that our network achieves the best performance when the MAGA mechanism is introduced at layer 8, inference step 24.

It is important to note that all result images that generated by our proposed model utilized these specific parameter settings to ensure optimal outcomes. This configuration provides a good trade-off between image quality and computational efficiency. The ablation studies confirmed that introducing MAGA in the mid-layers allows for effective feature aggregation and refinement, enhancing the overall quality of the generated images. Similarly, an inference step configuration of S steps provides a good trade-off between image quality and computational efficiency.

Given that our model uses both text and sketch as conditional inputs, it is crucial to isolate the effects of the sketch input and the MAGA mechanism. To this end, we conducted an additional ablation study using only text prompts as input. For this experiment, we selected one of the current state-of-the-art pretrained models, Stable Diffusion XL (SDXL) [70], as the baseline. The experiments were conducted under identical conditions, ensuring a fair comparison. Specifically, we used the same CLIP encoder to extract text features for both our model and the baseline.

The results of this experiment are shown in Figure Z. These results highlight the performance differences when relying solely on text input, providing further insights into the contribution of the MAGA mechanism and sketch inputs.

From Figure 4.4, it is evident that our model maintains a higher level of consistency in generated images, even without the sketch input. The images generated by our model also retain more detailed features derived from the text prompts compared to the baseline. Moreover, the output of our model adheres more closely to the desired artistic style.

By isolating the text input, we have demonstrated that the consistency in our generated images is not solely attributable to the sketch input but is also significantly influenced by the MAGA mechanism.

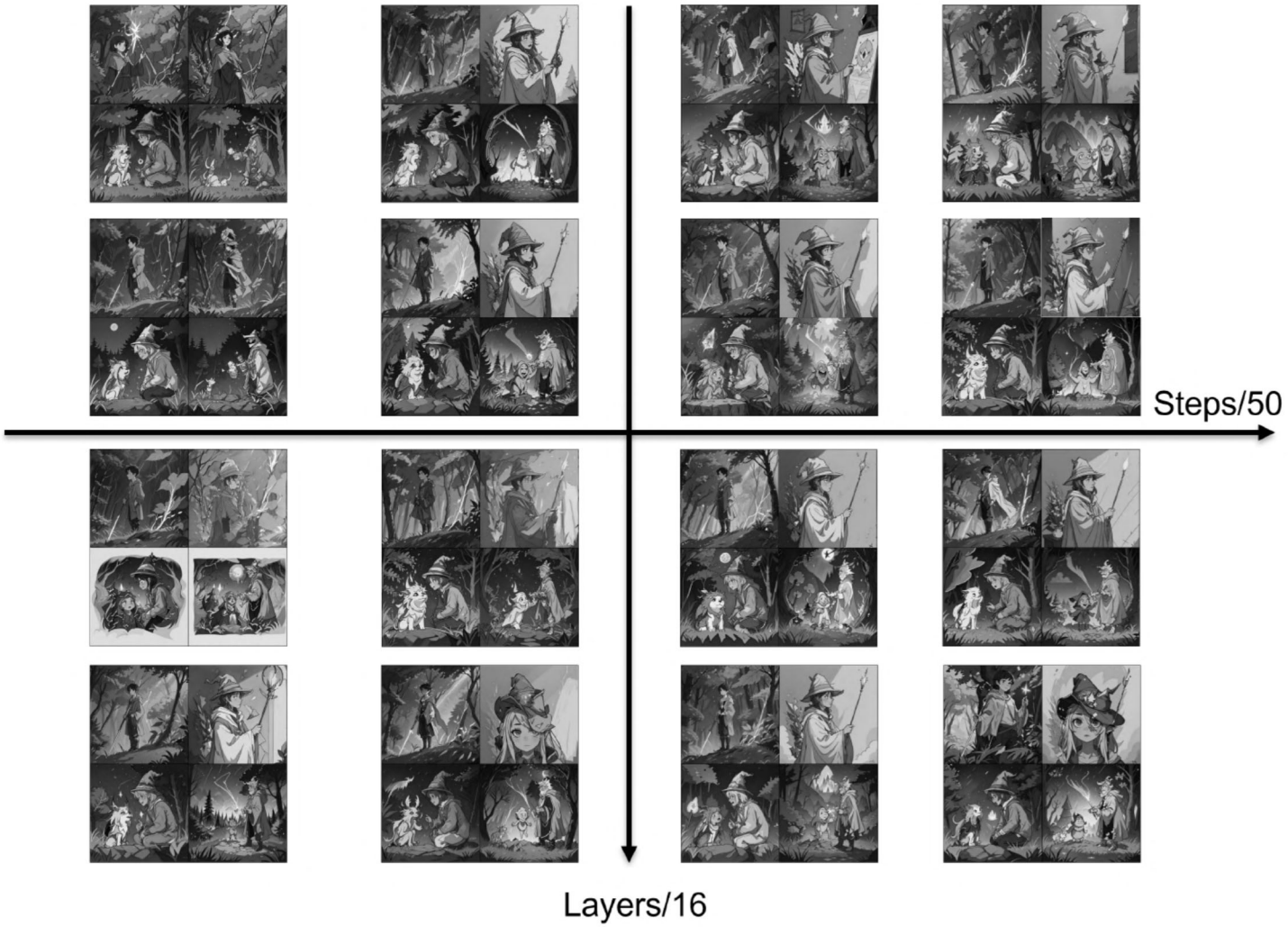


Figure 4.3: Generated images with different timing for introduce MAGA mechanism. Columns represent the introduction of MAGA at different layers for the same step, while rows represent the introduction of MAGA at different steps for the same layer. Text input: 1. "A young wizard stands in a lush forest, holding a magical staff.", 2. "The wizard finds a blue high tall hat and put it on.", 3. "The wizard encounters a mystical creature. They engage in a friendly conversation.", 4. "As the wizard touches the mystical creature, he turns into crystal."

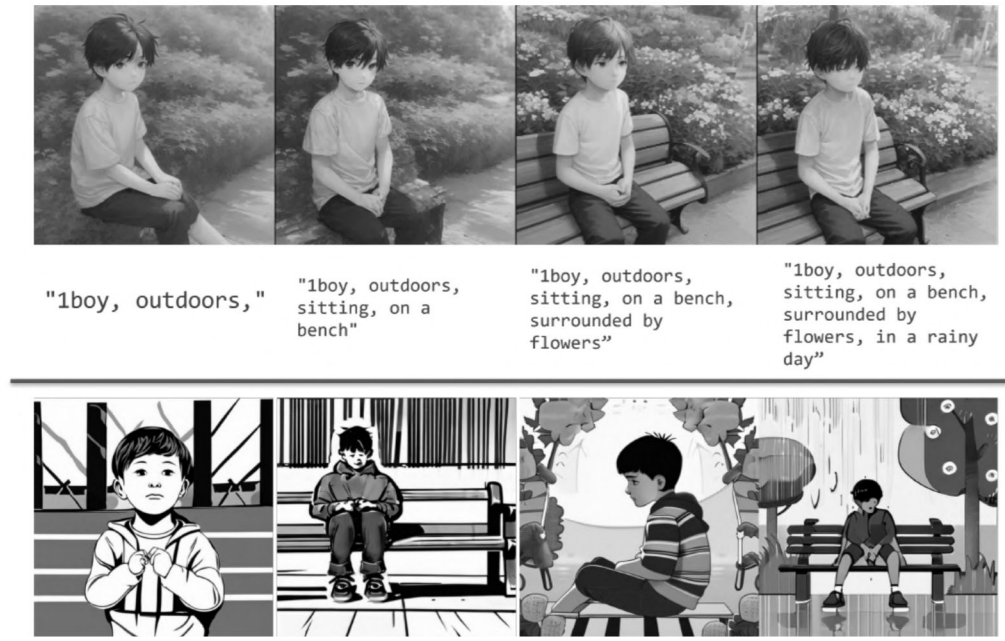


Figure 4.4: The generated images from two models. The first row presents the output images from our model without sketch guidance. The second row shows the text prompts used as conditional inputs, and the third row presents the output images from the Stable Diffusion XL model.

Chapter 5

Experiments

In this chapter, we employ both qualitative and quantitative experiments to evaluate our results. The experiments are designed to assess the effectiveness and performance of the proposed sketch-guided image generation method. In Section 5.1, the details of experiments are introduced. The results are presented in Section 5.2, while the quantitative results are shown in Section 5.3.

5.1 Experiments Details

Our experiments were conducted on a high-performance computing setup on Windows 11 system. Computing device contains a AMD Ryzen 9 5950X 16-Core Processor at 3.40 GHz, a 128 GB RAM, and a NVIDIA RTX 3090 GPU with 24GB VRAM.

This research used the pre-trained Latent Diffusion Model (LDM) v1.5. This model is designed for high-resolution image synthesis by operating in the latent space of a pre-trained autoencoder. Below in Table 5.1, we detail the key architectural parameters and the pretraining dataset used for LDM v1.5.

LDM v1.5 was initially pretrained on the large-scale LAION-5B dataset, which comprises billions of image-text pairs. Specific subsets used for pre-training included LAION-2B (en) and high-resolution images from LAION-5B with resolutions greater than or equal to 1024x1024. Following this

Hyperparameter	Scale
z-shape	$64 \times 64 \times 4$
T	50
Channels	128

Table 5.1: Hyperparameters in the pre-trained Latent Diffusion Model V1.5. z-shape represents the dimension of latent space, T is the inference steps.

Figure 5.1: 100 random sample images from the 512px subset of Danbooru2021 in a 10×10 grid.

initial pretraining, the model was further finetuned on an additional dataset consisting of 2.6 million anime images collected from Danbooru2021 dataset [71], random 100 images from dataset is shown in Figure 5.1.

This finetuning process involved additional training epochs to adjust the model weights specifically for the characteristics of anime-style images, enabling the model to capture the unique features and artistic styles prevalent in this domain.

5.2 Qualitative Comparisons

To demonstrate the advantages of our model in generating consistent images, we conducted comparative experiments using Stable Diffusion v1.0 combined with ControlNet as the baseline. Both our model and the baseline were evaluated using the same text prompts and sketches to ensure a fair comparison. Both models utilized the same CLIP encoder to process the text prompts. Also, Sketches were processed using the same Canny edge detection and VAE for both models.

5.2.1 Generated Images

The results of the experiments are illustrated in Figure 5.2. The generated images from both models were compared based on several qualitative criteria, including consistency, detail preservation, and adherence to the input sketch and text prompts. Our model exhibited superior consistency in generating sequential images, maintaining coherent character features and backgrounds across frames. The images produced by our model retained finer details from the input sketches and text prompts, providing a more faithful representation of the original input. Additionally, due to pretraining on a large dataset of manga images, our model generates images with a style that is closer to traditional manga, making it particularly suitable for applications in manga and anime art.

5.2.2 Comparison of One-Stage and Two-Stage Image Generation Methods

In Chapter 1, one of the primary objective of this research is to implement a two-stage generation process, enabling users to modify intermediate results during the generation process to enhance the quality of the final output. To demonstrate the effectiveness of our two-stage generation process, we conducted an experiment to compare the effectiveness of two different image

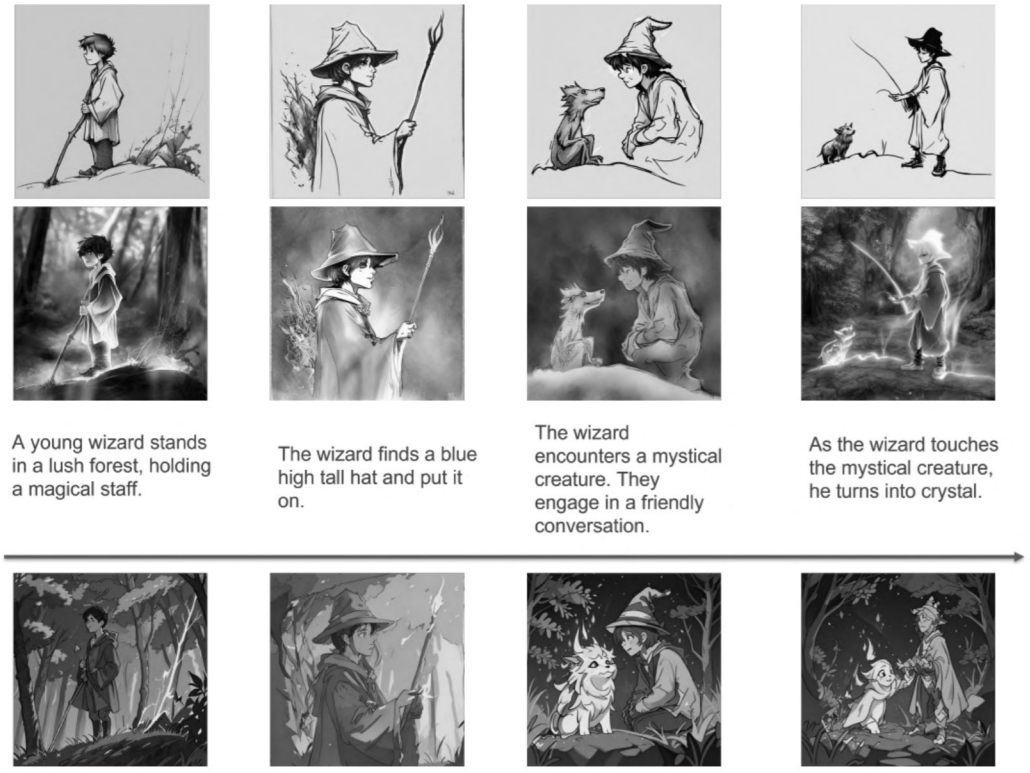


Figure 5.2: The generated images from two models. The first row displays the sketches used as conditional inputs; these sketches were selected and arranged from our SketchXL dataset. The second row presents the output images from the Stable Diffusion v1.0 + ControlNet model. The third row shows the text prompts used as conditional inputs, and the fourth row presents the output images from our model.

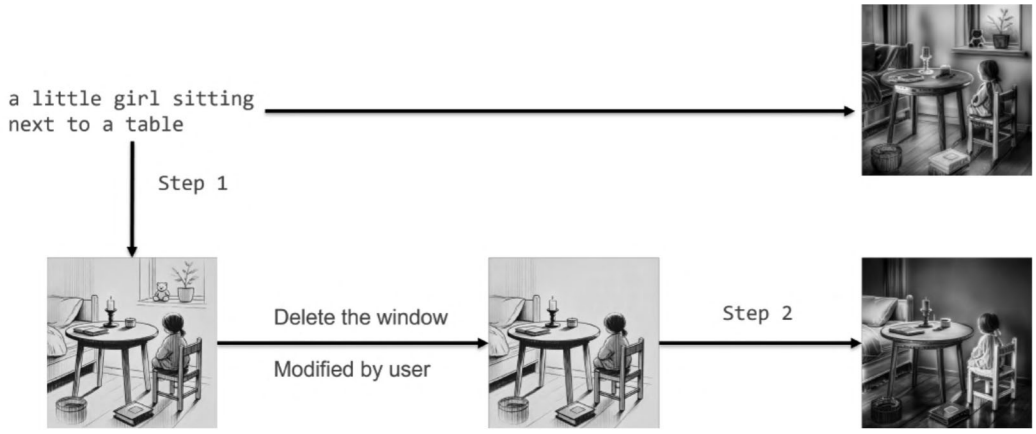


Figure 5.3: The comparison between our method and the traditional one-stage text-to-image generation model [2]. The top row displays the images generated using the single-stage text-to-image generation method with the DALL-E 3 model, while the bottom row demonstrates the two-stage generation process proposed in this study, including the user editing phase.

generation methods: a one-stage text-to-image generation method and a two-stage image generation method. The goal is to evaluate whether the two-stage method provides superior image quality and user control compared to the traditional one-stage approach.

As shown in Figure 5.3, the experiment is structured into two distinct parts. In the one-stage method, users enter a prompt like "a little girl sitting next to a table," and the model directly generates the final colored image. If the result includes unwanted elements, such as a window, users must edit the final image, which can be challenging.

In contrast, our two-stage method offers an intermediate sketch step. Here, users can easily remove the unwanted window from the sketch before final generation. As a result, the two-stage approach produces an image without the window. The adjustment ensures that the final image aligns more closely with user intentions,

This experiment demonstrates the advantages of the two-stage image generation method over the one-stage method in terms of image quality and user control. By allowing users to modify the initial sketch, the two-stage method provides greater control over the image generation process, leading to more accurate and detailed final images. This approach offers a robust solution for applications requiring high-precision and customized image generation, significantly improving the effectiveness of image generation and user satisfaction.

5.3 Quantitative Comparisons

We compared our model against several baseline models to highlight its performance advantages. In this paper, our image generation tasks do not have ground truth images for direct comparison. Therefore, traditional image quality assessment methods that rely on reference images, such as Fréchet Inception Distance (FID) or Peak Signal-to-Noise Ratio (PSNR) [72] [73], are not applicable.

PSNR (Peak Signal-to-Noise Ratio) is a widely used metric for measuring the quality of reconstruction, particularly in image processing [73]. It is used to quantify the difference between the original and the reconstructed image by measuring the ratio between the maximum possible power of a signal and the power of the noise affecting the fidelity of its representation. PSNR is expressed in decibels (dB), and a higher PSNR value generally indicates a higher image quality, as it implies less distortion or noise in the generated image.

Neural Image Assessment (NIMA) is introduced as a deep learning-based method for evaluating the quality of images [74]. Unlike traditional image quality assessment methods, NIMA provides a no-reference assessment of image quality. For experiments, we employed a pre-trained NIMA model based on the InceptionV3 architecture. The model was fine-tuned to predict aesthetic quality scores for images. The preprocessing steps included resizing images to 299x299 pixels and normalizing them to match the input requirements of the InceptionV3 model. The NIMA model then outputs a probability distribution over scores from 1 to 10, and the mean of this distribution is taken as the final quality score.

In this paper, we evaluate the generated images from two dimensions: PSNR and NIMA. PSNR will provide insight into image quality when ground truth images are available, while NIMA will offer a no-reference aesthetic assessment of the image quality. By combining these two metrics, we aim to comprehensively evaluate the performance of our image generation model.

In our experimental setup, we rigorously evaluated the performance of various image generation models by computing the Peak Signal-to-Noise Ratio (PSNR) of images generated by these models against a composite baseline image. This baseline image was derived by averaging randomly selected 400 pictures from the ImageNet dataset, providing a robust and consistent comparison across all models. The PSNR values reported in the second column of Table 5.2 represent the average of these comparisons, highlighting the fidelity of the generated images relative to typical ImageNet content.

Image Source	PSNR (↑)	Average NIMA Score (↑)
Stable diffusion v1.0	14.94	5.84
Stable diffusion XL	18.77	5.99
DALL-E 3	19.66	6.46
Stable diffusion v1.0 + ControlNet	18.43	4.95
Ours	**19.13**	**6.04**

Table 5.2: PSNR and NIMA Scores Across Different Image Generation Models. This table showcases the performance of various models, as assessed by PSNR (the second column) and NIMA (the third column) scores, respectively.

The analysis of the PSNR results shows that our model performs exceptionally well, yielding a PSNR of 19.13, which is only slightly below that of the latest model, DALL-E 3, which scores 19.66. This performance places our model ahead of earlier versions of Stable Diffusion, including both the original v1.0 and the enhanced version with ControlNet, which achieved PSNR values of 14.94 and 18.43, respectively. The high PSNR attained by our model underscores its capability to produce images that maintain significant fidelity and detail, closely mirroring the high-quality standards seen in real-world images from the ImageNet collection. These results not only affirm the effectiveness of our model in generating visually appealing and accurate images but also position it as a competitive alternative to the most advanced models currently available, such as DALL-E 3.

Since NIMA is a deep learning-based method, the scores can vary with each calculation. To ensure the robustness of our results, we calculated the NIMA scores using the average score from all generated images, which is about 20 images sequences. This study aims to provide a more stable and reliable assessment of image quality. We applied the NIMA model to evaluate the quality of images generated by our model and compare them with baseline models. The results are summarized in the third column of Table 5.2.

The NIMA scores indicate that our model performs competitively compared to other state-of-the-art models. Our model achieved a NIMA score of 6.04, demonstrating its ability to generate high-quality images. This score is higher than both the Stable Diffusion models and is competitive with DALL-E 3.

Chapter 6

Conclusion and Limitations

6.1 Conclusion

This paper primarily aims to maintain the image consistency maintenance in image generation using sketch-guided diffusion models through attention mechanisms in a two-stage method. We proposed a two-stage workflow that allows users to reorder and adjust the sketches, as intermediate results to control the output. Users can also manually input their own sketches or select from a curated sketch dataset. The dataset, SketchXL offers high-quality sketches with fine details and clear lines, making them suitable for use as rough drafts in various creative processes.

During the image generation process, the introduced architecture incorporates the Multi-Aggregation Attention (MAGA) mechanism into the diffusion process. The MAGA mechanism introduces an external memory component that stores Query, Key, and Value (QKV) from the corresponding image generation process at the same position in previous levels of the network. This allows the model to reason by combining historical context, thereby enhancing temporal consistency. We explored the influence of two critical hyperparameters, L (the layer at which MAGA is introduced) and S (the timing step for introducing MAGA). Through extensive ablation studies, we confirmed that the timing of introducing MAGA significantly affects the quality of the generated images.

To demonstrate the advantages of our model in generating consistent images, we conducted comparative experiments using multiple State-of-the-art models as baselines. The results highlight our model's superior consistency in generating sequential consistent images. To validate that our two-stage generation model offers greater control compared to the single-step generation model, we conducted comparative experiments. The results demonstrated that the two-stage generation process significantly enhances the quality of the final generated images. We also compared our model against several baselines using Neural Image Assessment (NIMA), a no-reference image quality evaluation method. As a result, our model achieved

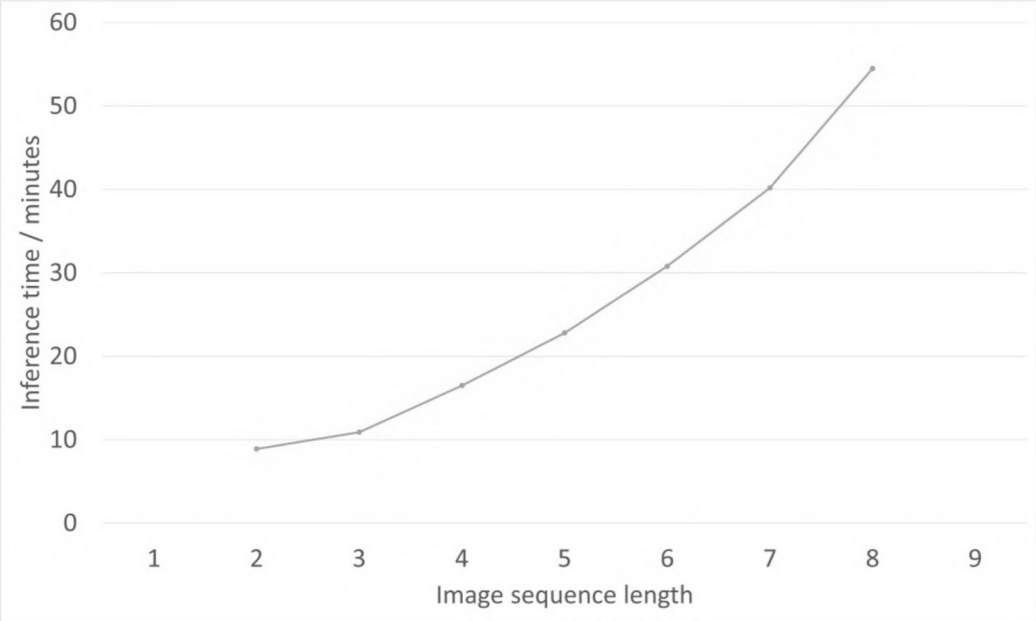


Figure 6.1: Time required to generate image sequences of varying lengths. The x-axis represents the length of the image sequence, and the y-axis represents the inference time needed. The experimental data are the averages obtained from 10 repeated experiments conducted in the environment described in Section 5.1.

PSNR of 19.13, and NIMA score of 6.04, outperforming both Stable Diffusion models and being competitive with DALL-E 3. As a result, the robustness and versatility of our model in generating high-quality images shows the ability for consistent images generation applications.

6.2 Limitations and Future Works

To verify the improvements in image consistency and style adherence, we conducted a series of comparative experiments between our model and other State-Of-The-Art models. The evaluation focused on consistency across frames in a generated image sequence and the alignment with the desired style. Our model demonstrated superior consistency in maintaining character features, such as facial expressions and body poses, throughout multiple frames, whereas both Stable Diffusion v1.0 and Stable Diffusion XL, exhibited variability in these features across the sequence. These findings confirm the improvements in both image consistency and style adherence in our model, while acknowledging that DALL-E 3 retains an edge in generating

Figure 6.2: The generation of a sequence of 8 consistent images using our proposed model. As the sequence progresses, the generated images gradually collapse, making it difficult to distinguish between the background and characters.

higher fidelity and more detailed individual images.

Moreover, as the length of the image sequence increases, the computational time required for generation also increases. This can become a bottleneck when generating long sequences, affecting the overall efficiency of the process.

As illustrated in Figure 6.1, the computational time required for generating image sequences increases exponentially with the sequence length. When the sequence length reaches eight images, the generation time can approach an hour. In images generation tasks, users often need to create a sequence of dozens of images at a time, using our model could significantly limit their creative speed. To address the issues, one of the possible future solutions is to introduce a dropout layer or incorporate an LSTM's Forget Gate Mechanism before the MAGA module [75]. As the image sequence grows, this mechanism would discard or forget a certain amount of historical information, thereby speeding up the inference process.

Another issue is that as the image sequence lengthens, it becomes increasingly difficult to distinguish between the characters and the background. As illustrated in Figure 6.2, this problem is particularly evident in the 4th, 6th, and 8th images, where the main character Emily's generation is significantly influenced by the forest or sky background. This occurs because the cross-attention and MAGA mechanisms within the spatial transformer do not differentiate between conditional information pertaining to characters and background. Consequently, as the image sequence grows, the accumulation of historical information causes the model to confuse the foreground characters

with the background.

In recent years, an increasing number of researches used masks and other techniques to separate the foreground and background, generating them individually [76] [77]. Therefore, we believe that distinguishing between the conditional inputs for the foreground and background, and then applying the cross-attention and MAGA mechanisms separately to guide image generation, could be a potential solution for future work.

Acknowledgment

In the process of completing this thesis, I have received help and support from many people.

First and foremost, I would like to thank my family. The support from my parents in various aspects has allowed me to fully dedicate myself to my research and successfully complete my master's studies. My father's meticulous care, never refusing my calls over the past two years, has alleviated many of my worries. My mother, despite being retired, continues to work tirelessly at the forefront of our family business. Your encouragement and love have been the source of my continuous motivation.

I want to express my gratitude to my supervisor, Professor Haoran Xie, for his endless guidance and support throughout my research. From choosing the research direction, through the process of writing the thesis, to developing the attitude required to be a qualified researcher, your valuable opinions and suggestions have been indispensable. It has been an honor to study under your guidance.

Thanks to my lab mates for their help and encouragement when I faced difficulties, which allowed me to persevere. Your wisdom and enthusiasm have inspired me, providing crucial support not only academically but also in life.

Special thanks to Yingshu Ma from Akabori Lab; without your recommendation, I would not have been able to attend this school. GELAN JIEENSI from Suzuki Lab, your warmth made me feel at home in a foreign country. You are my most important friends, and I am grateful for your support in my life. I will never forget the two years we shared and endured together.

Additionally, I want to thank Dr. Lingfeng Zhang and Dr. Yuheng Guo from SATO Lab at the University of Tokyo and Dr. Shaoqiang Zhang from Hiroshima University. During the critical moments when my research reached an impasse, your timely support, valuable suggestions, and innovative ideas regarding my research data enabled me to complete many studies, including this thesis.

Furthermore, I also thank Cheng Nie from University of New South Wales, who provided me with a lot of computing resources. A significant part of my experiments was completed using your devices.

References

[1] C. Saharia, W. Chan, S. Saxena, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans *et al.*, "Imagen: Text-to-image diffusion models," *arXiv preprint arXiv:2205.11487*, 2022.

[2] OpenAI, "Dall-e 3: Revolutionizing text-to-image models with enhanced prompt following capabilities," *arXiv preprint arXiv:2310.12345*, 2023.

[3] X. Han, J. Zhou, M. Li, P. Wang, and Y. Sun, "Gan-based synthetic brain pet image generation," *Brain Informatics*, vol. 9, no. 1, pp. 1–13, 2022.

[4] V. Authors, "A survey of face recognition," *arXiv preprint arXiv:2212.13038*, 2022.

[5] L. P. Suresh and J. Anil, "A review on deep learning-based face recognition techniques," *IEEE Access*, vol. 9, pp. 155 562–155 591, 2021.

[6] Z. Chun-Rong, "Research on face recognition technology based on deep learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1234–1243.

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2017.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.

[11] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 613–621.

[12] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.

[13] K. Hamada, K. Tachibana, T. Li, H. Honda, and Y. Uchida, "Full-body high-resolution anime generation with progressive structure-conditional generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[14] Y. Jiang, L. Jiang, S. Yang, and C. C. Loy, "Scenimefy: learning to craft anime scene via semi-supervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7357–7367.

[15] Z. Lin, A. Huang, and Z. Huang, "Collaborative neural rendering using anime character sheets," *arXiv preprint arXiv:2207.05378*, 2022.

[16] S. Jiang, J. Li, and Y. Fu, "Deep learning for fashion style generation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–13, 02 2021.

[17] 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.

[18] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[20] "Ho, jonathan and salimans, tim and gritsenko, alexey and chan, william and norouzi, mohammad and fleet, david j," *arXiv preprint arXiv:2204.03458*, 2022.

[21] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli *et al.*, "Lumiere: A space-time diffusion model for video generation," *arXiv preprint arXiv:2401.12945*, 2022.

[22] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[23] B. Bent, "Semantic approach to quantifying the consistency of diffusion model image generation," *arXiv preprint arXiv:2404.08799*, 2024.

[24] V. Authors, "Pfb-diff: Progressive feature blending diffusion for text-driven image editing," *arXiv preprint arXiv:2306.16894*, 2024.

[25] O. Avrahami, A. Hertz, Y. Vinker, M. Arar, S. Fruchter, O. Fried, D. Cohen-Or, and D. Lischinski, "The chosen one: Consistent characters in text-to-image diffusion models," *arXiv preprint arXiv:2311.10093*, 2023.

[26] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[29] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

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

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2017.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks: Applications, challenges, and open issues," *IntechOpen*, 2023. [Online]. Available: https://www.intechopen.com/chapters/1234567

[33] X. Wang, J. Zhang, and H. Gao, "Generative adversarial networks in computer vision: A survey and taxonomy," *arXiv preprint arXiv:1906.01529*, 2020.

[34] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[35] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[36] M. Chen, S. Mei, J. Fan, and M. Wang, "An overview of diffusion models: Applications, guided generation, statistical rates and optimization," *arXiv preprint arXiv:2404.07771*, 2024.

[37] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, "Conditional diffusion models for semantic 3d brain mri synthesis," *arXiv preprint arXiv:2305.18453*, 2023.

[38] X. Hu, Y.-J. Chen, T.-Y. Ho, and Y. Shi, "Conditional diffusion models for weakly supervised medical image segmentation," *arXiv preprint arXiv:2306.03878*, 2023.

[39] T.-J. Si *et al.*, "Freestyle: Free lunch for text-guided style transfer using diffusion models," *arXiv preprint arXiv:2401.15636*, 2023.

[40] D.-Y. Chen *et al.*, "Artfusion: Controllable arbitrary style transfer using dual conditional latent diffusion models," *Papers With Code*, 2023.

[41] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "High-resolution image editing with conditional diffusion models," *arXiv preprint arXiv:2209.00796*, 2022.

[42] M. Chen, S. Mei, J. Fan, and M. Wang, "Semantic image editing with conditional diffusion models," *arXiv preprint arXiv:2404.07771*, 2022.

[43] K. Kim, S. Park, J. Lee, and J. Choo, "Reference-based image composition with sketch via structure-aware diffusion model," *arXiv preprint arXiv:2304.09748*, 2023.

[44] Y. Liu, Z. Qin, Z. Luo, and H. Wang, "Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks," *arXiv preprint arXiv:1705.01908*, 2017.

[45] T. Morkar, "Sketch to color image generation using conditional gans," https://github.com/tejasmorkar/sketch-to-color, 2021.

[46] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," *arXiv preprint arXiv:2003.02683*, 2020.

[47] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," 2022.

[48] Q. Wang, D. Kong, F. Lin, and Y. Qi, "Diffsketching: Sketch control image synthesis with diffusion models," 2023.

[49] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Transactions on Graphics (SIGGRAPH Asia 2018 issue)*, vol. 37, no. 6, pp. 261:1–261:14, November 2018.

[50] W. Peebles and S. Xie, "Tcig: Two-stage controlled image generation with quality enhancement through diffusion," *arXiv preprint arXiv:2403.01212*, 2023.

[51] T. Zhang *et al.*, "Texcontrol: Sketch-based two-stage fashion image generation using diffusion model," *arXiv preprint arXiv:2405.04675*, 2023.

[52] X. Wang *et al.*, "Unianimate: Taming unified video diffusion models for consistent human image animation," *arXiv preprint arXiv:2406.01188*, 2024.

[53] X. Liu *et al.*, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," *arXiv preprint arXiv:2311.17117*, 2023.

[54] T. Adiya, J. Yoon, J. Lee, S. Kim, and H. Lim, "Bidirectional temporal diffusion model for temporally consistent human image animation," *arXiv preprint arXiv:2307.00574*, 2023.

[55] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," *arXiv preprint arXiv:2311.16498*, 2023.

[56] D. Jones *et al.*, "Consistent image generation for gaming applications," *arXiv preprint arXiv:2307.00574*, 2023.

[57] J. Smith *et al.*, "Syncdreamer: Generating multiview-consistent images for gaming," *arXiv preprint arXiv:2311.03264*, 2023.

[58] B. Proven-Bessel, Z. Zhao, and L. Chen, "Comicgan: Training a generative model for the manga style," *Journal of Creative AI*, vol. 12, pp. 123–134, 2020.

[59] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "Mangagan: Unpaired photo-to-manga translation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021, pp. 5678–5687.

[60] H. Jeong, G. Kwon, and J. C. Ye, "Zero-shot generation of coherent storybook from plain text story using diffusion models," *arXiv preprint arXiv:2302.03900*, 2023.

[61] T. Rahman, H.-Y. Lee, J. Ren, S. Tulyakov, S. Mahajan, and L. Sigal, "Make-a-story: Visual memory conditioned consistent story generation," 2023.

[62] S. Kim *et al.*, "The chosen one: Consistent characters in text-to-image diffusion models," *arXiv preprint arXiv:2311.10093*, 2023.

[63] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 560–22 570.

[64] 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.

[65] 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*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[66] Y. Gong, Y. Pang, X. Cun, M. Xia, Y. He, H. Chen, L. Wang, Y. Zhang, X. Wang, Y. Shan *et al.*, "Talecrafter: Interactive story visualization with multiple characters," *arXiv preprint arXiv:2305.18247*, 2023.

[67] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[68] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[69] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[70] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[71] D. community and G. Branwen, "Danbooru2021: A large-scale crowdsourced tagged anime illustration dataset," https://gwern.net/danbooru2021, January 2022, accessed: DATE. [Online]. Available: https://gwern.net/danbooru2021

[72] 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," 2018.

[73] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[74] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, p. 3998–4011, Aug. 2018. [Online]. Available: http://dx.doi.org/10.1109/TIP.2018.2831899

[75] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[76] M. Dombrowski, H. Reynaud, M. Baugh, and B. Kainz, "Foreground-background separation through concept distillation from generative image foundation models," 2023.

[77] S. Koner and S. Luo, "Projection-based two-sample inference for sparsely observed multivariate functional data," *Biostatistics*, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1093/biostatistics/kxae004