Master's Thesis

Real-Time Drawing Assistance with Sketch-Based Control via Diffusion Models

CHEN Chuang

Supervisor Haoran Xie

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

September, 2025

Abstract

Creating high-quality illustrations presents a significant challenge for beginners, primarily due to the high demands for stylistic complexity, fine-grained detail, and proportional accuracy inherent in artistic creation. Traditional drawing assistance systems, such as retrieval-based approaches, typically rely on pre-constructed databases, lack flexibility, and struggle to effectively accommodate diverse user inputs. In this context, this study proposes an efficient interactive drawing assistance system capable of generating multiple guidance sketches in real time to support user drawing. The system also allows users to select from various drawing styles, including anime, realistic, and others, based on their preferences. It not only lowers the creative threshold for beginners, helping them complete their illustrations more smoothly, but also provides rapid and high-quality support for users with basic drawing skills, thereby improving overall drawing efficiency and creative freedom.

The system is built upon the StreamDiffusion inference framework, an optimized diffusion-based architecture capable of generating high-quality RGB images in real time. To meet the demands of various artistic styles, the system adopts Low-Rank Adaptation to efficiently fine-tune the Stable Diffusion model, achieving rapid style adaptation with only a small number of style-specific samples. The system supports multimodal input, allowing users to simultaneously provide hand-drawn sketches and textual prompts, thus enabling the generation of RGB images consistent with the selected style. These generated images are subsequently processed by the Informative Drawings model, which extracts structural features by integrating geometric and semantic information, producing rough sketches. To further enhance the effectiveness of the sketches as drawing guidance, a custom-designed denoising optimizer is implemented, which employs image filtering techniques to remove unnecessary noise while preserving essential edge structures, ultimately producing clean and structurally coherent guidance sketches.

To systematically evaluate the usability and effectiveness of the proposed system, a user study was conducted involving participants with varying levels of drawing experience. The experiment compared three types of drawing interfaces: (1) a baseline interface without any assistance, (2) a shadow-guidance interface with a reference sketch fixed beneath the canvas, and (3) the proposed drawing assistance interface. The usability of the proposed interface was quantitatively assessed using the System Usability Scale (SUS), and user preferences were collected to statistically analyze how effectively each interface supported users in achieving their intended drawings.

To further verify the system's effectiveness during the actual drawing process, expert and non-expert reviewers were invited to evaluate the users' drawings. Evaluation criteria included overall shape accuracy, local proportion consistency, and line clarity and expressiveness. Experimental results demonstrate that the proposed system significantly improves drawing quality. The SUS scores also reflect users' high satisfaction with the system's ease of use and overall experience. Preference analysis further indicates that, compared to other interfaces, the proposed system enables users to produce drawings that better align with their original creative intentions, highlighting its effectiveness in supporting personalized artistic expression.

In summary, this study presents a real-time drawing assistance system that integrates the StreamDiffusion image generation framework, LoRA fine-tuning, a structural sketch extraction model, and a filter-based denoising optimization module. The system provides efficient and accurate drawing guidance across a range of artistic styles, significantly enhancing users' drawing efficiency and creative experience, and demonstrating strong potential for practical application and broader adoption.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research background and the key challenges faced by current drawing assistance systems in supporting beginners. Section 1.1 highlights the importance of drawing as a form of cultural and artistic expression and discusses the difficulties beginners encounter in mastering proportions, rendering details, and applying stylistic elements, thereby underscoring the need for dedicated assistance tools. Section 1.2 analyzes the limitations of existing systems, such as their strong dependence on pre-constructed datasets. Section 1.3 defines the primary objective of this study, which is to develop a real-time drawing assistance system that supports multiple artistic styles while maintaining a low barrier to entry. Section 1.4 summarizes the main contributions of this research, including the development of a real-time drawing assistance system capable of supporting multiple artistic styles, the integration of diffusion models for high-quality image generation, and the optimization of interaction design to improve accessibility for beginners.Finally, Section 1.5 outlines the overall structure of this thesis.

1.1 Research Background

Drawing is one of the most significant forms of human artistic and cultural expression, having played a central role since ancient times in conveying emotions, recording stories, and showcasing creativity. Both traditional drawing and digital illustration are widely applied across various fields, including entertainment, gaming, product design, and digital media. Despite its strong artistic appeal and cultural value, and its ability to attract numerous enthusiasts, creating high-quality artworks remains a highly challenging task for beginners. The creation of high-quality artworks typically requires mastery of complex skills such as proportion, composition, detail rendering, and stylis-

tic expression. As illustrated in Figure 1.1, these skills are usually acquired through long-term training and professional guidance, which poses a significant barrier for many art enthusiasts attempting to transform their creative ideas into complete works. While beginners are often driven by the passion to produce works that meet their own expectations, their final pieces frequently fall short, as they struggle to accurately express the images they envision. Over time, this can lead to a loss of confidence and even abandonment of artistic pursuits.

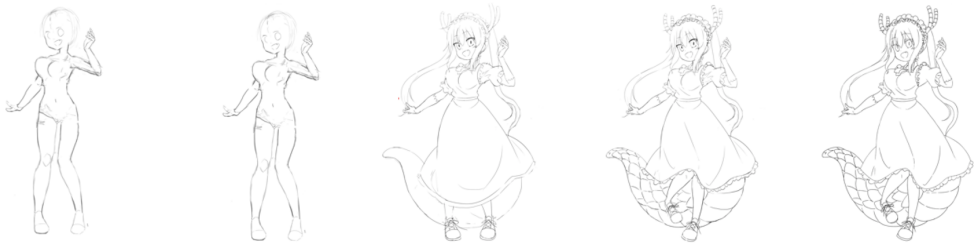


Figure 1.1: An example of a manual drawing process illustrating the progressive refinement from a rough body outline to a complete character with detailed costume and accessories. This highlights the complexity of traditional illustration, which typically requires advanced skills and step-by-step construction. Source: Adapted from `https://note.com/manyplay_aihara/n/n79c8359c4b54`.

Beginners need to acquire fundamental knowledge of human body proportions, facial structure, expression depiction, and the use of light and shadow. Professional artists typically undergo years of systematic training to apply these skills flexibly in their creative work. For beginners, even with some theoretical understanding, it is often difficult to accurately capture the symmetry and details of facial features, such as the size and placement of the eyes, the height of the nose bridge, or the curvature of the mouth. Minor inaccuracies in these elements can make a character appear unbalanced. Furthermore, depicting vivid expressions, such as the gaze during a smile or the forehead lines when frowning, presents additional challenges. This requires keen observation of facial muscle movements as well as the ability to express them through precise lines and appropriate contrasts of light and shadow. Individuals of different age groups have distinct characteristics, such as the roundness of a child's face or the wrinkles and looseness of an elderly person's skin. In addition, varying drawing styles, including realism, manga, cartoon, or abstract, demand different techniques, all of which require long-term prac-

tice to master gradually. Faced with the difficulty of fully expressing their inner creativity, many beginners gradually lose confidence and may even choose to give up.

Traditional drawing tools, such as Photoshop [1], Clip Studio Paint [2], and Procreate [3], offer powerful functionalities for artistic creation. However, these software applications are primarily designed for professional users with advanced skill levels and typically feature numerous complex functions and interfaces. Although they enable the production of high-quality artworks, they require substantial time and effort to master. For beginners, these tools do not directly reduce the difficulty of drawing; instead, they may impose additional cognitive and operational burdens. This highlights the importance of developing intuitive, user-friendly drawing assistance systems that support multiple artistic styles. The emergence of drawing assistance systems offers new opportunities for users without access to professional guidance, helping them overcome technical barriers and enhancing their creative experience.

In the study of drawing assistance systems, early approaches primarily relied on data retrieval techniques, dynamically matching users' hand-drawn sketches with similar samples from a database to provide reference and inspiration. For example, the representative work PortraitSketch [4] allows users to upload images and combine them with line tracing to enhance the aesthetic quality of their sketches, thereby supporting the creation of more visually appealing works. However, such methods heavily depend on the scale and diversity of the underlying database and place considerable demands on the user's own drawing skills. As a result, they exhibit clear limitations in terms of creative flexibility and applicability.

With the rapid advancement of generative models, researchers have progressively introduced them into the domain of drawing guidance. A representative example is Pix2Pix [5], a conditional generative adversarial network (GAN) [6] capable of transforming input sketches into realistic images, which has been widely applied in graphic design and creative illustration tasks. Another example is Sketch2Model [7], a method that enables the conversion of sketches into three-dimensional models, allowing users to generate corresponding 3D representations from hand-drawn sketches, thereby offering new tools for engineering design and 3D modeling. In addition, Sketch-RNN [8], a generative approach based on recurrent neural networks, learns the sequential features of sketches, making it possible to automatically generate and complete sketches, and demonstrating promising applications in sketch automation. A previous study [9] proposed a novel drawing assistance framework based on diffusion model, which allows users to select a desired drawing style and then generates multiple guidance sketches in that specific style. The system also improves user interface interaction to enhance usability.

Although generative models offer expanded possibilities for the functionality of drawing assistance systems, they still face several challenges in practical applications. For example, Pix2Pix is highly sensitive to the details and quality of the input sketches, requiring users to have a certain level of drawing proficiency. While Sketch-RNN enables sketch generation, its outputs are often simplified and struggle to meet the demands of complex or detailed creations. Moreover, most of these models rely on specific training datasets, resulting in limited flexibility when adapting to multiple styles and diverse scenarios.

In recent years, the rapid development of image generation technologies has offered new opportunities for the advancement of drawing assistance systems. In particular, text-to-image generation methods based on diffusion models [10], such as MidJourney [11] and DALL-E [12], have demonstrated remarkable generative capabilities across various fields, including artistic creation, visual design, and game development. These models produce high-quality images through a progressive denoising process, requiring only simple textual prompts to generate diverse and detailed results. This technology has attracted widespread attention from professional creators and has increasingly permeated the practices of general users and enthusiasts, enabling individuals without professional drawing skills to obtain artworks that meet their expectations.

1.2 Research Issues

Early drawing assistance systems primarily relied on dataset-driven approaches, with their functionalities typically built upon large-scale, pre-constructed sketch or image libraries. These systems commonly adopted retrieval-based methods to provide users with drawing references. Although such approaches offer a certain degree of assistance, they also present notable limitations. On the one hand, the content generated by the system is highly dependent on the size and diversity of the dataset, which limits the flexibility in representing style, subject matter, and details. Methods based on GAN model exhibit similar issues. As shown in Figure 1.2, ISketchNFill [13] offers a limited number of assistance categories, making it difficult for the system to provide effective guidance when users' creative needs go beyond the scope of the dataset. This limitation significantly hinders the realization of diverse and personalized creations. On the other hand, the interaction design of current drawing assistance systems also remains suboptimal. As shown in Figure 1.3, which illustrates the SketchHelper [14] drawing assistance interface, the system displays multiple guidance sketches in a stacked format, resulting in a

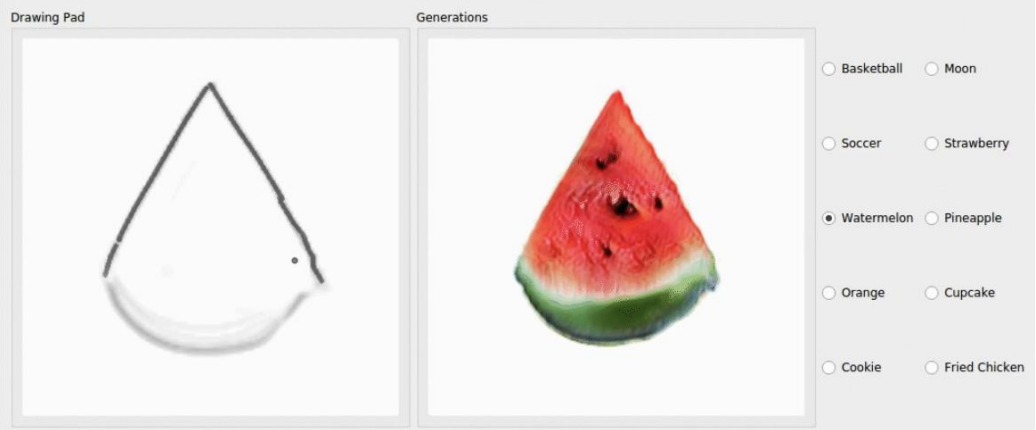


Figure 1.2: Illustration of the ISketchNFill drawing assistance interface. The left panel shows the drawing pad where users input sketches. The center panel displays the generated image conditioned on the selected category and the user's drawing. The right panel provides a list of predefined drawing categories (e.g., watermelon, cupcake) for users to choose from. Source: Adapted from `https://github.com/arnabgho/iSketchNFill?tab=readme-ov-file`

visually cluttered and less navigable interface. Such a presentation makes it challenging for users to quickly identify and select a suitable guidance sketch that aligns with their creative intent, thereby reducing both the usability and overall user experience of the system.

Diffusion models, with their progressive denoising generation mechanism, have demonstrated strong potential for high-quality image synthesis, offering new technological avenues for drawing assistance systems. However, such models are typically computationally intensive and slow in generation, creating a tension with the requirements of drawing assistance systems for real-time performance and smooth feedback. In practical applications, balancing consistency in generation with real-time responsiveness to ensure continuous and immediate user interaction during the creative process remains a critical technical challenge.

In addition, the design of interactive experiences [15, 16] remains a major shortcoming of current drawing assistance systems. Many deep learning-based tools are complex to operate and contain numerous parameters, making them difficult to learn and use for individuals without professional backgrounds. Designing intuitive and user-friendly interfaces that allow users to quickly become proficient and receive timely, effective guidance during the creative process is essential for achieving system usability and broad appli-

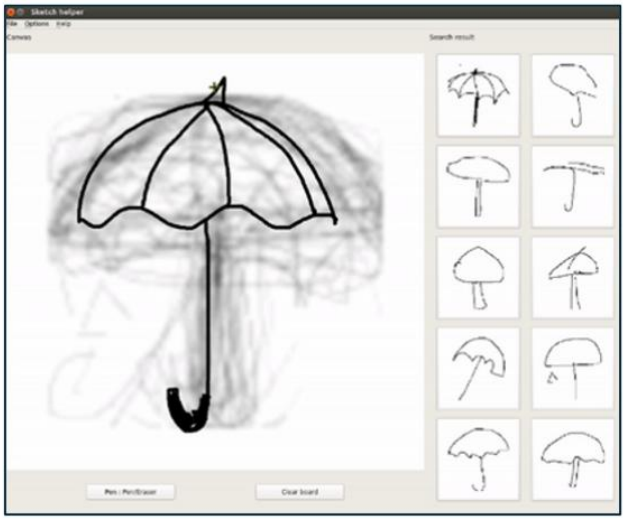


Figure 1.3: Illustration of the SketchHelper drawing assistance interface. The left canvas displays shadowed stroke guidance based on retrieved candidates, while the right panel shows the top-10 similar stroke-based sketch suggestions. The system supports step-by-step sketch retrieval and real-time stroke guidance for structured sketch construction. Source: Adapted from `https://github.com/kookmin-hci/sketch_helper?tab=readme-ov-file`

cation.

1.3 Research Objective

This study is centered around the following key questions: How can a drawing assistance system be designed to support multiple artistic styles, thereby meeting beginners' flexible needs across diverse creative contexts? How can advanced diffusion models be integrated to enhance the system's responsiveness to user intent and improve the precision of local detail representation? And how can a low-barrier, highly user-friendly interactive interface be developed to enable beginners to interact smoothly with the system and receive meaningful support and feedback during the creative process? Specifically, the main objectives of this study can be summarized as follows:

First, this study explores how to construct a system framework based on diffusion models to support multi-style drawing generation, enabling the system to flexibly meet diverse creative needs. Unlike traditional drawing tools limited to a single style, the proposed system aims to provide users with a wide range of artistic style options, such as anime and realism, helping

beginners explore and learn across multiple styles, thereby expanding their artistic expressiveness. The overview of the proposed method in this study is shown in Figure 1.4

Second, the study focuses on designing a low-barrier, highly user-friendly interactive interface that is not only suitable for creators with professional backgrounds but also offers beginners an intuitive and accessible user experience. This objective involves optimizing interface layout, feature presentation, and operational workflows, enabling users to interact efficiently with the system without complex learning, and to receive real-time feedback and effective drawing guidance. In summary, this research constructs a drawing assistance system with both academic significance and practical application potential, supporting beginners in better achieving their creative goals and advancing new explorations of human–computer collaboration in digital art creation.

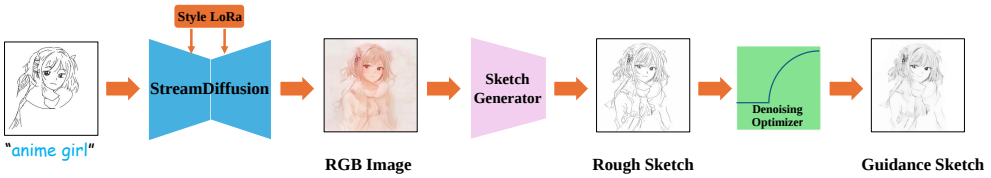


Figure 1.4: The overview of proposed drawing assistance system. The proposed system takes a hand-drawn sketch and a text prompt as inputs. StreamDiffusion first generates an image based on these inputs. This image is then passed to the Sketch Generator, which produces a rough sketch. To enhance clarity, the rough sketch is further refined by the denoising optimizer, yielding a clean guidance sketch tailored for multi-style illustration.

1.4 Originality and Significance

This study makes several theoretical and practical contributions to the design and implementation of an interactive drawing assistance system for beginners, summarized as follows:

(1) Proposal of a multi-style, multi-context drawing assistance framework. Existing drawing assistance tools are often limited to specific styles or tasks. The system developed in this study supports a wide range of artistic styles, including anime, realism, and fantasy, enabling users to explore and practice across styles during learning and creation. By offering diverse style options

and flexible application scenarios, the system significantly expands the applicability of drawing assistance tools, providing beginners with greater creative space and learning resources.

(2) Optimized interaction design to lower the barrier for beginners. The study emphasizes beginner-oriented interaction design, focusing on interface simplicity, operational intuitiveness, and user-friendliness. By incorporating guided operations, visual feedback, and an easy-to-understand feature layout, the system helps reduce the entry barrier for beginners and provides continuous positive support throughout the creative process, enhancing their learning motivation and creative confidence.

(3) Systematic user evaluation and application validation. To comprehensively assess the system's effectiveness and practicality, the study designs and conducts a systematic user study involving both quantitative and qualitative evaluations. By collecting feedback from users with diverse backgrounds (including beginners and experienced artists), the study analyzes the system's performance in multi-style support, detail control, and interaction experience. This not only verifies the system's effectiveness in practical applications but also offers valuable insights for the design and optimization of similar systems.

In conclusion, this study achieves innovative advances across multiple aspects, including algorithm integration, system design, interaction optimization, and user validation. It enriches the research perspectives of interactive drawing assistance and provides new directions and practical cases for exploring the application of generative artificial intelligence in creative tools.

1.5 Outline of Thesis

The structure of this paper is constructed as follows:

Chapter 1 introduces the research background and the key challenges faced by current drawing assistance systems in assisting beginners. It highlights the importance of developing efficient and controllable generative tools for conceptual architectural exploration. The chapter concludes by outlining the research objectives, major contributions, and the overall organization of this thesis.

Chapter 2 reviews related works relevant to the proposed framework, focusing on neural network architectures and diffusion-based generative models. The chapter discuss diffusion model evolution, including DDPM, DDIM, LDM, and the efficiency-oriented LCM-LoRA. Finally, it summarizes previous research on drawing assistance systems.

Chapter 3 introduces the core components of the proposed drawing as-

sistance system. It begins with Stable Diffusion for high-quality image generation, followed by the StreamDiffusion inference framework designed for real-time image synthesis. The chapter concludes with the Informative Drawings model, which extracts semantically meaningful sketch representations to guide user drawings.

Chapter 4 presents the guidance sketch generative framework designed to assist users in improving drawing efficiency and creative quality. The chapter explains the end-to-end process—from user input to RGB image generation, sketch extraction, and optimization—and describes how different styles are supported via Style-LoRA. It also discusses the system's interface design and acceleration strategies for real-time interaction.

Chapter 5 presents the user study evaluating the system's usability, including participant demographics, experimental design, and questionnaires for collecting subjective feedback, and provides an objective evaluation of the drawing outcomes through expert and non-expert assessments based on criteria such as shape accuracy, proportion, and line quality.

Chapter 6 presents the results of both subjective and objective evaluations. It reports system usability scores, drawing performance metrics, and qualitative feedback from participants. The chapter also includes comparative results across three interfaces to demonstrate the advantages of the proposed system.

Chapter 7 summarizes the research findings and discusses future directions. It highlights the effectiveness of the proposed system in guiding users' drawing processes and improving creative outcomes, identifies current limitations, and proposes future enhancements such as local editing capabilities and the integration of more advanced conditional generation techniques.

Chapter 2

Related Works

This chapter reviews related works closely associated with the proposed framework, with a focus on generative models. Section 2.1 provides a systematic overview of the evolution of diffusion models, ranging from the foundational Denoising Diffusion Probabilistic Models to the accelerated Denoising Diffusion Implicit Models and the latent-space-based Latent Diffusion Model, and further introduces the Latent Consistency Model [17] as a solution to improve inference efficiency. Section 2.2 focuses on Low-Rank Adaptation [18] and its application in our system, particularly the integration of Low-Rank Adaptation with Latent Consistency Model to enable efficient and high-quality image generation. Section 2.3 introduces previous related research

2.1 Generative Models

This section provides a systematic overview of the development and key advancements in diffusion models. Section 2.1.1 introduces the fundamentals of Denoising Diffusion Probabilistic Models [19], including their modeling process. Section 2.1.2 discusses the contribution of Denoising Diffusion Implicit Models [20] in accelerating the sampling procedure. Section 2.1.3 discusses the application of Variational Autoencoders in latent space modeling and image compression. Section 2.1.4 focuses on Latent Diffusion Models [21], which leverage latent space representations to enable high-resolution image synthesis with reduced computational cost. Finally, Section 2.1.5 presents the Latent Consistency Model [17], which improves inference efficiency through consistency learning and serves as a foundation for the integration of LCM-LoRA [22] in subsequent sections.

2.1.1 Denoising Diffusion Probabilistic Models

Jonathan Ho et al. proposed Denoising Diffusion Probabilistic Models (DDPM), a deep generative framework inspired by nonequilibrium thermodynamics. The central idea of DDPM is to construct a Markovian forward process that progressively adds Gaussian noise to the original data x_0 following a predefined variance schedule β_t, ultimately transforming the data into nearly pure noise x_T. Figure 2.1 shows the denoising and denoising process of ddpm. Specifically, the forward process is defined as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

and its joint distribution can be written as:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}).$$

To recover data samples from noise, the model defines a corresponding reverse process:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

and employs variational inference to optimize its parameters. Unlike conventional approaches, DDPM introduces a simplified objective that directly predicts the noise component ϵ added at each step, formulated as:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right],$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

This design not only simplifies training but also establishes connections between DDPM, denoising score matching across multiple noise scales, and annealed Langevin dynamics. Furthermore, the authors highlight that the iterative denoising process of DDPM can be interpreted as a generalized form of progressive decoding, offering greater flexibility compared to traditional autoregressive generation.

In experimental evaluations, DDPM achieved remarkable performance on benchmarks such as CIFAR-10 [23], LSUN [24] and CelebA-HQ [25]. For instance, on CIFAR-10, the model reached an Inception Score of 9.46 and an FID score of 3.17, surpassing many state-of-the-art GAN-based models at the time. Overall, this work not only laid a solid theoretical foundation for diffusion models but also significantly influenced the development of subsequent methods, including DDIM, Latent Diffusion, Stable Diffusion [21], and Imagen [26].

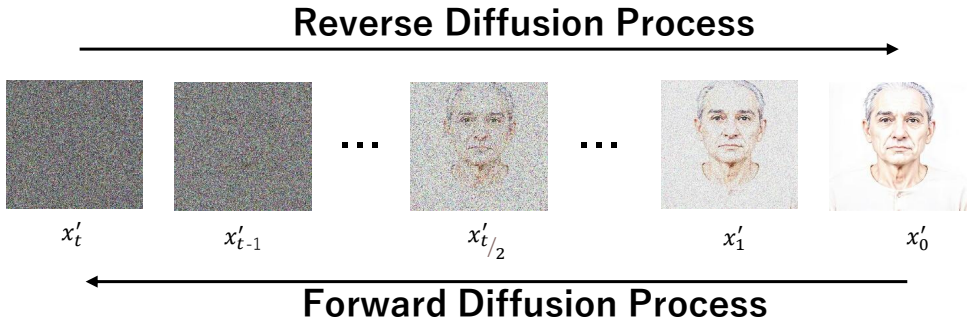


Figure 2.1: In the forward diffusion process, noise is progressively added to a clean image over multiple time steps until it becomes pure noise. The reverse diffusion process then reconstructs the original data by gradually removing this noise. By learning this denoising trajectory, the model is able to generate new data samples from random noise.

2.1.2 Denoising Diffusion Implicit Models

While DDPM have demonstrated remarkable performance in high-quality image generation, they suffer from a significant drawback: the sampling process requires simulating a lengthy Markov chain, typically with over a thousand steps, gradually denoising from Gaussian noise to the data distribution. This results in extremely slow sampling, making DDPM impractical for tasks requiring efficiency or real-time generation. To address this limitation, Song et al. proposed Denoising Diffusion Implicit Models (DDIM), an effective extension of DDPM that achieves substantial acceleration by introducing a non-Markovian inference process, all while maintaining the same training objective as DDPM.

The key insight of DDIM lies in the observation that, although DDPM relies on a stepwise Markovian noise addition process during training, its objective only depends on the marginal distributions $q(x_t|x_0)$ at each timestep, not on the full joint distribution $q(x_{1:T}|x_0)$. Leveraging this fact, DDIM formulates a non-Markovian inference process:

$$q_\sigma(x_{1:T}|x_0) := q_\sigma(x_T|x_0) \prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0),$$

with the conditional distributions defined as:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I\right),$$

and when $\sigma_t = 0$, the model reduces to a deterministic process without stochastic noise sampling.

The corresponding sampling update rule is:

$$x_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(x_t) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t} f_\theta(x_t)}{\sqrt{1 - \alpha_t}} + \sigma_t \epsilon,$$

where $f_\theta(x_t) = (x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t))/\sqrt{\alpha_t}$, and $\epsilon \sim \mathcal{N}(0, I)$.

Compared to DDPM, DDIM offers several unique advantages. First, it can generate high-quality samples using significantly fewer steps (e.g., 20 to 100 steps instead of the standard 1000), achieving 10 to 50 times acceleration without retraining the model, and even outperforming DDPM in some settings. Second, due to its deterministic nature, DDIM enables semantically meaningful interpolation and smooth transitions in the latent space, a feature difficult to achieve with stochastic samplers like DDPM. Furthermore, DDIM can serve as an encoder, mapping observations to latent codes and deterministically reconstructing them, demonstrating latent space modeling capabilities that are uncommon in conventional diffusion models. Figures 2.2 clearly show the difference between ddim and ddpm.

In summary, DDIM is an important advancement over DDPM, not only significantly improving sampling efficiency but also expanding the practical and theoretical scope of diffusion models, paving the way for future work in conditional generation, controllable generation, and latent space manipulation.

2.1.3 VAE

Variational Autoencoders (VAE) [27] is a probabilistic generative model composed of an encoder and a decoder. The encoder maps input data to the parameters of a latent distribution (typically the mean and variance), from which a latent vector is sampled. The decoder then reconstructs the input from this latent representation. Unlike traditional autoencoders, VAE employs variational inference by maximizing the Evidence Lower Bound, which approximates the true posterior and regularizes the latent space. This is commonly achieved by minimizing the Kullback-Leibler divergence between the learned latent distribution and a prior, often a standard Gaussian. Such regularization promotes smooth, continuous, and semantically meaningful latent representations, enabling diverse and coherent generation. In diffusion models, VAE is frequently used to embed high-dimensional images into a lower-dimensional latent space, reducing computational cost and enabling efficient latent-space denoising. This makes VAE particularly suitable for high-resolution image synthesis tasks.

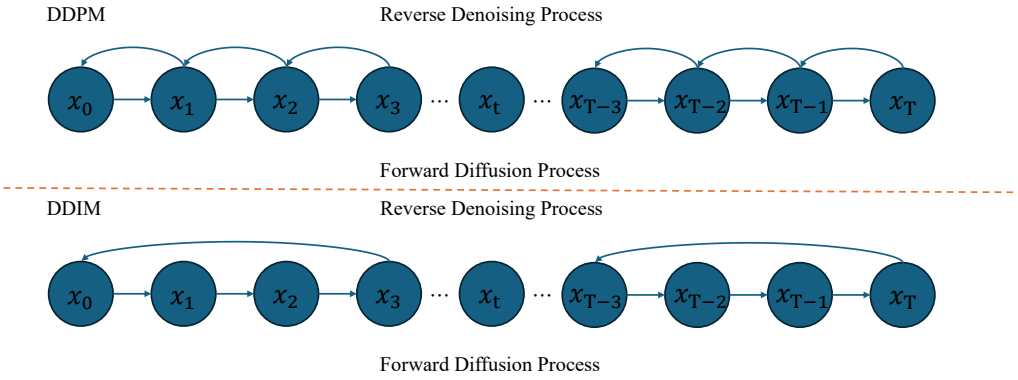


Figure 2.2: Comparison between DDPM and DDIM in the reverse denoising process. DDPM performs denoising sequentially through all timesteps, recovering the image from x_T to x_0 step by step. In contrast, DDIM enables a non-Markovian sampling process that allows skipping intermediate steps, thus reconstructing x_0 more efficiently from fewer steps, such as directly from x_3.

2.1.4 Latent Diffusion Model

Although diffusion models [28, 29, 30] have demonstrated strong capabilities in image generation, performing the denoising process directly in pixel space requires substantial computational resources and memory during inference, especially for high-resolution images. This limitation has hindered their practical deployment in real-time or resource-constrained scenarios. To address this challenge, the Latent Diffusion Model (LDM) was introduced, with the core idea of shifting the diffusion process from the pixel space to a compressed latent space. Compared to pixel-space diffusion models, LDM significantly reduces computational complexity while maintaining excellent performance in preserving image details. This balance between reduced complexity and preserved detail enables high visual fidelity in the generated results. Figure 2.3 shows the core method of LDM.

This design significantly reduces computational and memory costs, particularly for high-resolution image synthesis. The LDM employs an image autoencoder to project images from the pixel space into a latent space representation, thereby achieving dimensionality reduction from a high-dimensional pixel space to a more compact latent space. Specifically, a raw image x is encoded into a latent representation $z = E(x)$ using an encoder E. The forward noise addition and iterative denoising processes are then conducted within the latent space. Upon completion of denoising, the resulting rep-

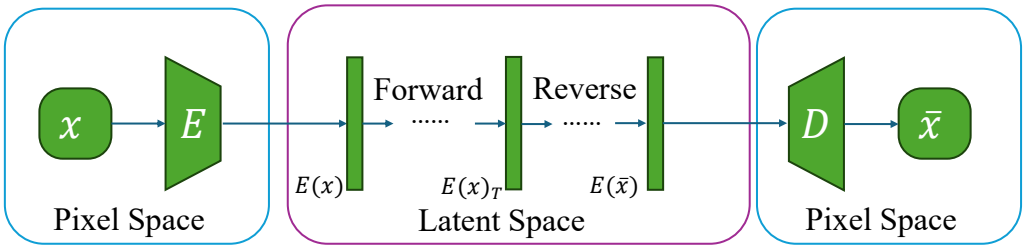


Figure 2.3: The process of LDM. The input image x is first encoded by an encoder E from the pixel space into a latent representation $E(x)$. A forward diffusion process is applied in the latent space, followed by a reverse denoising process that reconstructs the latent representation $E(\bar{x})$. The final latent is decoded by a decoder D to generate the reconstructed image $\bar{x}$ in the pixel space.

resentation is given by $\bar{z} = E(\bar{x})$. Finally, a decoder D reconstructs the image by mapping the latent representation back to the pixel space, namely $\bar{x} = D(\bar{z})$. Compared to conventional diffusion models operating directly in pixel space, LDM offers two notable advantages. First, by reducing the dimensionality of the diffusion process, it dramatically improves computational efficiency and reduces memory requirements, enabling high-resolution generation on standard hardware setups. Second, it becomes easier to incorporate textual inputs or other forms of conditional guidance in latent space, thereby enhancing the model's controllability and expressive capacity.

Empirical evaluations have demonstrated that LDM achieves strong performance across image generation, matching or even surpassing pixel-space diffusion models, while significantly reducing computational resources and improving inference speed. Performing diffusion operations in the latent space not only speeds up sampling but also helps to integrate other conditions to guide image generation, such as textual prompt. Importantly, LDM has laid the groundwork for large-scale conditional generative systems such as Stable Diffusion XL (SDXL) [31], driving the broader practical adoption of diffusion models across research and industry.

2.1.5 Latent Consistency Model

LDM compress high-dimensional information into a low-dimensional latent space, which reduces computing resources for image generation. However, inference still relies on multi-step iterative denoising, typically requiring 25 to 50 steps to obtain satisfactory results. This stepwise sampling leads to slow inference and limits the practicality of LDM in applications with strin-

gent real-time or low-latency requirements. Consistency Models (CMs) [32] can achieve fast sampling with few steps while maintaining generation quality, but prior studies have focused mainly on low-resolution image generation tasks such as ImageNet [33] 64×64 and LSUN 256×256, leaving their suitability for high-resolution synthesis insufficiently validated. To address the slow inference of LDM in high-resolution image generation, Luo et al. proposed Latent Consistency Models (LCMs) as a general acceleration framework applicable to pretrained LDM.

Similar to CMs, LCMs formulate the denoising process as solving an optimized ordinary differential equation. In contrast to the two-stage distillation used by Guided-Distill [34], LCMs adopt a one-stage guided distillation procedure that significantly reduces training time. LCMs directly learn to predict the PF-ODE solution in the latent space. To achieve this, they introduce a consistency loss during training that enforces agreement of latent representations across different time steps, thereby learning a stable mapping. As a result, LCMs substantially reduce the number of required sampling steps, thereby achieving fast and efficient inference process.

In practice, LCMs typically require only 2 to 4 steps to produce high-quality images. Under comparable step counts such as 1, 2, or 4 steps, they deliver superior image quality to prior methods, including DPM [35] and DPM++ [36], for 512×512 and 768×768 generation tasks. This makes LCMs particularly suitable for latency-sensitive applications such as interactive drawing, real-time video synthesis, and web-based image generation systems.

2.2 Low-Rank Adaptation

This section presents key techniques for improving adaptability and inference speed in diffusion models. Section 2.2.1 focuses on Low-Rank Adaptation (LoRA), which enables efficient fine-tuning by injecting trainable low-rank matrices into pre-trained models. Section 2.2.2 describes LCM-LoRA, which integrates LoRA with the Latent Consistency Model to enable high-quality image generation with minimal sampling steps, supporting real-time generation scenarios.

2.2.1 Low-Rank Adaptation

LoRA provides a way to efficiently fine-tune a pre-trained models with a large number of parameters. The core idea of LoRA is to freeze the weights of the pre-trained model and introduce two low-rank matrices for training

on a specific task. These low-rank matrices effectively capture the weight differences between the pre-trained model and those required for the specific task. After training, the adapted weights are integrated with the pre-trained model weights to achieve adaptation to the specific task. LoRA has two distinct advantages. First, it enables efficient fine-tuning with limited data. For instance, fine-tuning for image style transfer often requires only a few dozen examples to achieve satisfactory results. Second, the training process is highly efficient and significantly faster than conventional full fine-tuning approaches. Figure 2.4 illustrates the fundamental working principle of LoRA. Specifically, LoRA models the weight update as:

$$W' = W + BA,$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are low-rank trainable matrices, and $r \ll d$ denotes the rank. W is the pretrained model weights. Here, W' denotes the adapted weight matrix obtained by adding the low-rank update term BA to the frozen pre-trained weights W. During training, the pre-trained weight matrix W remains frozen, and only the low-rank matrices A and B are optimized to achieve task-specific parameter adaptation. As illustrated in Figure 2.4, the input vector $x \in \mathbb{R}^d$ is projected through both the frozen weight matrix W and the low-rank adaptation branch composed of A and B. Initially, the matrix A is randomly initialized from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, while B is set to zero. During fine-tuning, only the parameters of trainable matrices are adapted, allowing the adaptation without modifying the frozen weights. After training, the LoRA module can be directly inserted into the original pre-trained model to perform efficient fine-tuning, without altering the pre-trained model weights. This modular design enables the model to be efficiently adapted to new tasks while retaining the expressive power of the pre-trained model. LoRA achieves lightweight fine-tuning and personalization, making it possible to specialize large models for diverse applications without compromising their generalization performance.

LoRA has been successfully applied in various diffusion model fine-tuning scenarios, particularly in conditional diffusion models. By fine-tuning critical modules such as the U-Net using LoRA, it becomes possible to efficiently specialize the model for specific tasks, styles, or input conditions, while preserving the general-purpose capabilities of the base model. In this study, LoRA is integrated into the proposed painting assistance system to enable efficient, low-cost model adaptation across different drawing styles. This allows the system to provide users with guidance sketches in a variety of artistic styles.

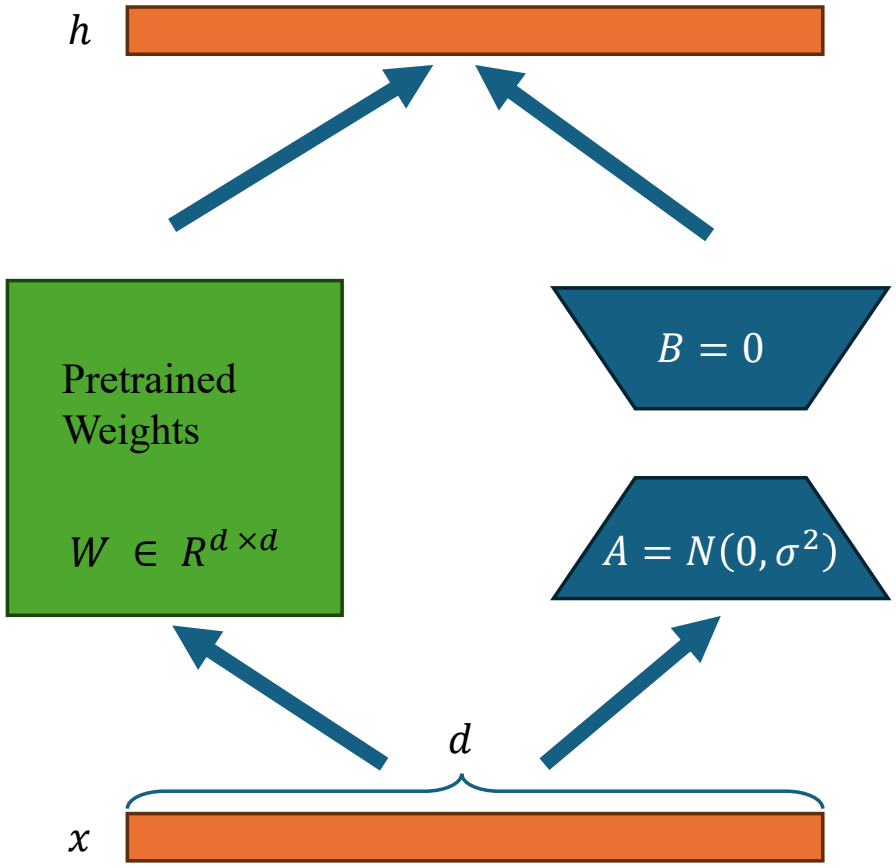


Figure 2.4: Illustration of the LoRA mechanism. The input features x are processed through both pretrained weights and a lightweight trainable low-rank adaptation module. The output h combines frozen knowledge and task-specific adaptation, enabling efficient fine-tuning with minimal parameter updates.

2.2.2 Latent Consistency Model with Low-Rank Adaptation

The LCMs has recently emerged as an effective framework for accelerating inference in diffusion models. The core idea of LCMs is to introduce consistency constraints within the latent space, allowing the model to achieve high-quality generation using only a few inference steps, without the need for the complete iterative denoising process typically required by diffusion models. By leveraging latent consistency guidance, LCMs significantly re-

duces inference time while maintaining generation fidelity, making it a key technique for improving the real-time performance of diffusion models.

Building on this foundation, LCM-LoRA integrates LoRA into the LCMs framework, combining the benefits of low-rank fine-tuning and accelerated latent-space inference. Specifically, LCM-LoRA introduces lightweight trainable parameters in the form of low-rank matrices into critical modules such as U-Net or Transformer blocks while freezing the main model weights. This design enables efficient style-specific adaptation with minimal additional parameter overhead. Compared to traditional full fine-tuning approaches, LCM-LoRA substantially reduces both computational cost and memory usage, offering greater flexibility to adapt across various inference scenarios and application requirements.

LCM-LoRA has been successfully applied in diverse diffusion-based generative tasks, including text-to-image synthesis, personalized style transfer, and real-time image editing. In this study, we incorporate LCM-LoRA into our painting assistance system to enable fast, efficient generation in response to user inputs such as sketches or text prompts. By combining the inference acceleration offered by LCM with the efficient adaptation provided by LoRA, the system achieves high-quality, multi-style generation with low latency, enhancing its usability and effectiveness in interactive human-AI co-creation.

2.3 Drawing Assistance Systems

Due to their intuitive and concise expressive nature, sketches play a vital role in fields such as content design and engineering modeling, enabling the rapid communication of creative concepts and structural ideas. In recent years, a range of sketch-based tools have been developed to expand the application scope of sketch input. For instance, Sketch2Model [7] enables the transformation of hand-drawn sketches into 3D models, providing users with an intuitive and efficient solution for 3D design. Similarly, Pix2Pix [5] converts sketches into photorealistic images and has been widely applied in creative design and image synthesis tasks. These approaches have demonstrated significant potential in improving design efficiency and enhancing visual expression capabilities.

In recent years, drawing assistance systems have emerged as important research directions in computer graphics and human-computer interaction. These systems [37, 13, 14] aim to lower the creative barrier for novice users, improve drawing efficiency, and provide intelligent feedback and guidance through algorithmic and interface design. Among early representative works, ShadowDraw [38] stands out as a pioneering system that combines image

retrieval and sketch. Figure 2.5 shows the ShadowDraw system interface diagram. ShadowDraw matches user sketches against a large-scale image database to retrieve edge features and provides "shadow"-like visual hints, guiding users toward more accurate sketching. The system integrates a real-time responsive interface, offering dynamic suggestions as the user draws, significantly enhancing the experience for beginners and non-experts.

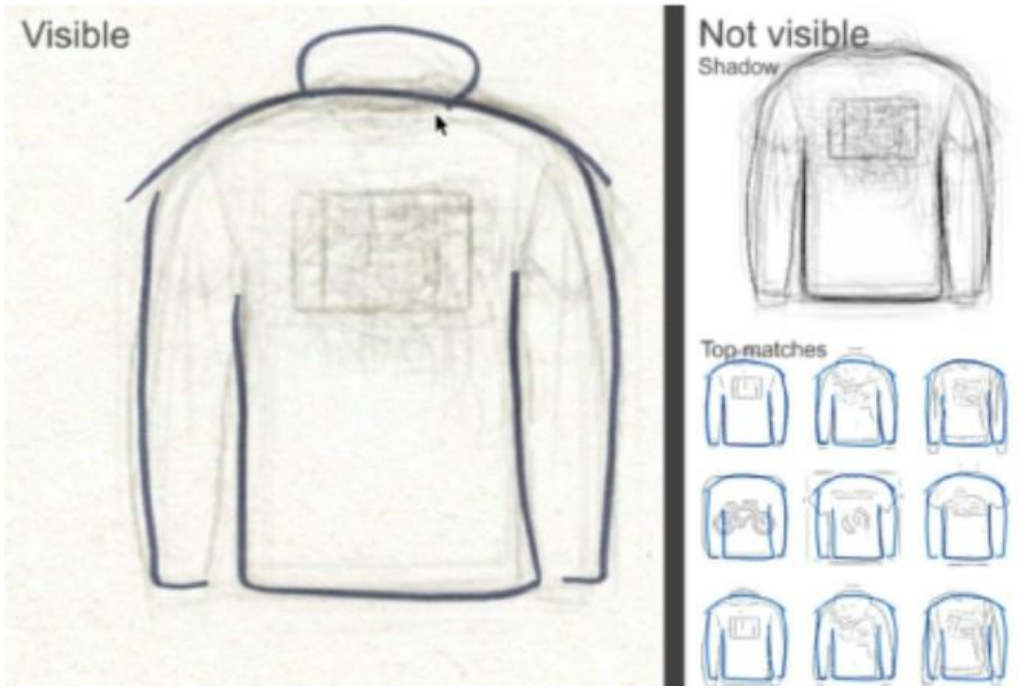


Figure 2.5: Illustration of the ShadowDraw user interface. The left panel shows the interface where users draw over a shadow guidance sketch. The right section, which is not part of the user interface, presents the retrieval results of the system. It displays the top nine images that are most similar to the user's input, along with the corresponding overlaid shadow guidance sketch. Source: Adapted from `https://www.youtube.com/watch?v=zh_-HUdQwow`

However, ShadowDraw also has notable limitations. Since it relies on database retrieval, it lacks generative capabilities, meaning that in creative scenarios beyond the scope of the dataset, the system offers little meaningful guidance. Additionally, the system emphasizes global outline suggestions but offers limited support for fine-grained, local-level adjustments.

Subsequent works [39, 40], such as Sketch-RNN [8] and DeepFaceDrawing [41], have introduced deep generative models to improve the diversity and in-

telligence of assistance tools. Sketch-RNN, built on a sequence-to-sequence recurrent neural network, learns stroke sequences and shape distributions to provide sketch generation and completion functionalities. However, its outputs are often constrained by the training distribution, limiting its adaptability to freeform creative inputs. DeepFaceDrawing integrates sketch-based interfaces with deep generative models to convert facial sketches into high-quality images, but its focus is largely confined to the domain of human faces, lacking generalizability across diverse categories and artistic styles. Furthermore, many of these systems offer fixed-step guidance in their interfaces, without multi-stage, progressive refinement, restricting users' ability to make localized, iterative improvements.

Previous studies have focused on enhancing the interactivity of drawing guidance systems, and others have explored improving their adaptability to a variety of artistic styles. For example, Dualface [42] introduced a two-stage interface for portrait sketching; Sketchhelper [14] integrates real-time stroke retrieval to provide stroke-level guidance; and EZ-Sketching employs a multi-level optimization strategy to automatically correct sketching errors. Moreover, studies have shown that interface features such as adjustable grids and variable canvas sizes can positively influence users' drawing performance. AniFaceDrawing [43] further proposes an unsupervised stroke-level decoupling training strategy to automatically align user sketches with the structural regions of anime-style faces. While these systems improve the user experience and assist in achieving better drawing outcomes, most of them are still limited in generalization ability, focus primarily on specific domains such as faces, and lack effective mechanisms for managing sketch clutter or supporting user-driven selection of guidance content.

These limitations highlight key challenges in existing drawing assistance systems, including limited generalization, lack of stylistic diversity, and inadequate support for multimodal inputs. Motivated by these challenges, this study integrates conditional diffusion models within a generative framework, paired with an interactive interface design. The proposed system offers real-time feedback and multi-style adaptability, aiming to provide flexible and efficient drawing guidance for users of varying skill levels.

Chapter 3

Preliminary Knowledge

This chapter introduces the key model components and inference framework employed in the proposed drawing assistance system, encompassing the overall workflow from high-quality image generation to semantic structure extraction. Specifically, Section 3.1 presents the foundational image generation model, Stable Diffusion, which enables efficient generation of high-resolution images by performing the diffusion process in latent space. Section 3.2 details the StreamDiffusion inference framework, which significantly reduces generation latency while preserving image quality, thereby meeting the system's requirements for real-time responsiveness and interactivity. Section 3.3 describes the Informative Drawings model [44], which integrates geometric and semantic constraints to extract informative line drawings from images, providing structurally clear sketch guidance to support users during the drawing process.

3.1 Stable Diffusion Model

Stable Diffusion [21] builds on the LDM method proposed by Rombach et al., and supports a variety of input conditions, such as images and semantic maps. It is one of the most widely adopted diffusion models currently. Compared to diffusion models operating directly in the pixel space, Stable Diffusion shifts the operations into the latent space. Stable Diffusion leverages a VAE to map the input image from the pixel space x into the latent space representation z. The forward diffusion process is then applied in the latent space to obtain z_t, followed by the denoising process to recover z.

Within the latent space, Stable Diffusion applies a forward noising process similar to the DDPM, where noise is gradually added to z_0 to produce a noisy

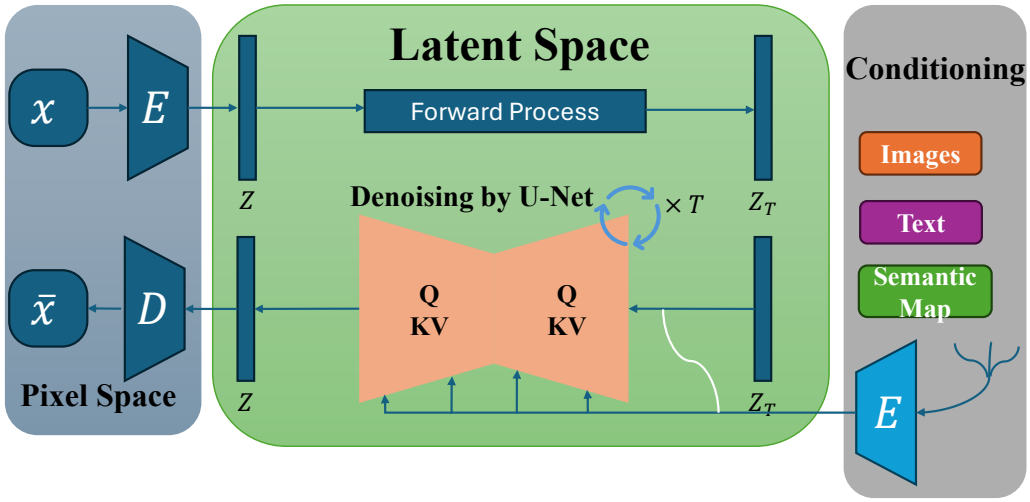


Figure 3.1: Overview of the Stable Diffusion architecture. The image encoder and decoder are components of a VAE, responsible for mapping images to and from a compressed latent space. The input image is first encoded into the latent space and perturbed with noise to simulate the forward diffusion process. A U-Net then performs iterative denoising within the latent space, guided by cross-attention with text embeddings generated by a pre-trained text encoder. Finally, the denoised latent representation is decoded by the VAE decoder to reconstruct the output image conditioned on the input text.

latent variable z_T. This process is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}),$$

where β_t denotes the noise schedule at time step t. The entire forward process is given by:

$$q(z_{1:T}|z_0) = \prod_{t=1}^{T} q(z_t|z_{t-1}).$$

During training, the model learns a parameterized noise predictor $\epsilon_\theta(z_t, t, c)$ that estimates the noise component added at each time step under conditional input c (typically, a text embedding). The objective is to minimize the following reconstruction loss:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right],$$

where $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

In the inference phase, the model starts from a noise sample $z_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it to recover the latent representation z_0, which is then decoded back to the image space using the VAE decoder D_ϕ:

$$\hat{x}_0 = D_\phi(z_0).$$

The architecture of Stable Diffusion consists of three main modules: a text encoder (typically the CLIP [45] text transformer) to encode textual prompt, a U-Net-based denoising network, and the VAE pair for latent image encoding and decoding. Figure 3.1 illustrates the architecture of Stable Diffusion. During the denoising process, cross-attention modules are integrated into the U-Net to inject conditional information from the text embeddings, thereby enabling controllable text-to-image generation.

In this study, Stable Diffusion serves as the fundamental generative module of our system. To enhance its adaptability to specific artistic styles, we incorporate a LoRA-based fine-tuning mechanism. Furthermore, to meet the stringent real-time requirements of interactive applications, we integrate the StreamDiffusion framework for inference acceleration, effectively reducing latency and improving user experience in interactive drawing guidance scenarios.

3.2 StreamDiffusion Inference Framework

StreamDiffusion [46] is a pipeline-level acceleration framework designed for real-time interactive generation using diffusion models. Figure 3.2 shows the

key framework of StreamDiffusion. Unlike prior efforts that focus on reducing the number of denoising steps or modifying the model architecture, StreamDiffusion restructures the inference process itself to optimize for throughput, latency, and energy efficiency. This framework incorporates three key components: Stream Batch Denoising, Residual Classifier-Free Guidance (R-CFG), and Stochastic Similarity Filter (SSF). In addition, Caching Mechanisms are integrated

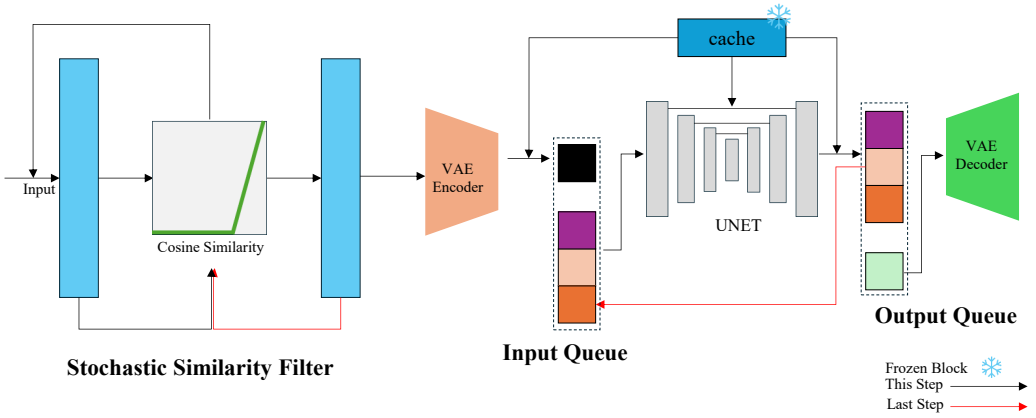


Figure 3.2: The StreamDiffusion framework incorporates three key components: Stochastic Similarity Filter, Stream Batch Denoising, and a multi-level caching strategy. Initially, the input image is evaluated by the Stochastic Similarity Filter to detect and discard frames that are overly similar to the previous ones, thereby eliminating redundant inputs. The filtered image is then decoded and added to the processing sequence. During inference, the framework employs prompt caching, noise caching, and scheduler caching to avoid redundant computations, significantly improving inference efficiency. Finally, the denoised latent representation is decoded into an RGB image for output.

In conventional diffusion models, denoising is performed sequentially, with inference latency scaling linearly with the number of steps. To address this, StreamDiffusion introduces the Stream Batch Denoising mechanism, which reformulates the denoising process to allow batch processing of multiple inputs across staggered timesteps. This design enables each forward pass of the U-Net to process several partially denoised images simultaneously, significantly reducing the per-frame inference time. Experimental results show that, under a 10-step denoising setting, Stream Batch Denoising achieves up to 2× acceleration compared to traditional step-wise inference.

To reduce the computational cost associated with Classifier-Free Guid-

ance (CFG) [47], StreamDiffusion further proposes R-CFG. Traditional CFG requires computing both positive and negative conditional predictions at each denoising step:

$$\epsilon_{\tau_i,\text{cfg}} = \epsilon_{\tau_i,\bar{c}} + \gamma(\epsilon_{\tau_i,c} - \epsilon_{\tau_i,\bar{c}}),$$

where γ is the guidance scale, and $\epsilon_{\tau_i,\bar{c}}$ is the noise predicted under the negative condition. R-CFG circumvents the repeated computation of $\epsilon_{\tau_i,\bar{c}}$ by approximating it using the initial latent x_0 and the current state x_{τ_i}:

$$\epsilon_{\tau_i,\bar{c}'} = \frac{x_{\tau_i} - \sqrt{\alpha_{\tau_i}} x_0}{\sqrt{\beta_{\tau_i}}},$$

$$\epsilon_{\tau_i,\text{cfg}} = \delta \cdot \epsilon_{\tau_i,\bar{c}'} + \gamma \left(\epsilon_{\tau_i,c} - \delta \cdot \epsilon_{\tau_i,\bar{c}'}\right),$$

where $\delta \in [0, 1]$ is a modulation factor controlling the magnitude of the virtual residual noise. This formulation effectively eliminates the need to evaluate the negative branch at each step, significantly accelerating multi-step inference.

To avoid unnecessary computations for static or near-identical inputs, StreamDiffusion employs a Stochastic Similarity Filter (SSF). This module computes the cosine similarity between the current input image I_{now} and the previous image I_{before}:

$$\text{Similar}(I_{now}, I_{before}) = \frac{I_{now} \cdot I_{before}}{\|I_{now}\| \|I_{before}\|},$$

$$P(\text{skip} \mid I_{now}, I_{\text{before}}) = \max\left\{0, \frac{\text{Similar}(I_{now}, I_{\text{before}}) - \lambda}{1 - \lambda}\right\},$$

where λ is a predefined similarity threshold. Based on this mechanism, the system may skip U-Net denoising, and VAE decoding for frames that exhibit minimal change. This technique dynamically reduces GPU utilization in static scenes and contributes to energy-efficient inference.

To accelerate inference, StreamDiffusion implements three caching strategies. The noise cache stores pre-sampled Gaussian noise tensors at each denoising timestep, enabling reuse across generations with identical settings and avoiding redundant sampling. The scheduler cache precomputes timestep values and scaling coefficients required during the denoising process, eliminating costly runtime lookups and improving speed, especially in real-time multi-frame scenarios. The prompt embedding cache retains the CLIP-encoded vectors of both positive and negative prompts, reducing repeated encoding costs. These caches collectively enhance the responsiveness and efficiency of the StreamDiffusion inference process.

Overall, StreamDiffusion provides a robust and efficient backend for real-time diffusion applications. Without compromising image quality, it achieves up to 91 frames per second on a single RTX 4090 GPU. Its combination of stream-aligned processing, lightweight guidance, and dynamic compute skipping makes it well-suited for deployment in interactive tasks such as sketch-based drawing systems, real-time video generation, and virtual avatars. Figure 3.3 shows some results generated by StreamDiffusion model.

3.3 Informative Drawings Model

The Informative Drawings model is a deep neural network architecture designed to generate line drawings that convey both geometric structure and semantic content. The core idea of this model is to regard line drawings as an encoded representation of images and to guide the generation process via explicit objectives related to geometry, semantics, and style. This enables the model to produce high-quality sketches from photographs without requiring paired training data.

The Informative Drawings adopts a GAN framework and follows an encoder-decoder architecture, where the encoder incorporates ResNet blocks for feature extraction. Several key loss functions are introduced during training to enhance the geometric accuracy and semantic consistency of the generated line drawings.

Specifically, the geometry loss supervises a pretrained depth estimation network to predict a depth map from the generated sketch, which is then compared to a pseudo-ground truth depth map of the input image. This encourages the model to place lines in geometrically meaningful locations, such as occluding contours. The semantic loss employs the CLIP image encoder to extract high-level features from both the input image and the generated sketch, enforcing semantic alignment by minimizing their cosine distance:

$$\mathcal{L}_{\text{semantic}} = \|\text{CLIP}(y) - \text{CLIP}(x)\|$$

In addition, the model incorporates a style consistency loss and a lightly weighted cycle-consistency loss to further enhance the naturalness and interpretability of the outputs. The full training objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CLIP}}\mathcal{L}_{\text{semantic}} + \lambda_{\text{geom}}\mathcal{L}_{\text{geometry}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}$$

In this study, the Informative Drawings model is employed as a preprocessing tool to extract structural sketches from RGB images. The generated

Figure 3.3: Results generated by StreamDiffusion. All four images were produced in "txt2img" mode using Stable Diffusion 1.5 model weights, denoised with four denoising steps, and generated at a resolution of 512×512. The first column shows results generated from the prompt "1 boy/girl with brown dog hair, thick glasses, smiling," while the second column corresponds to the prompt "a photorealistic portrait of a young man/woman."

sketches exhibit strong structure-preserving properties and semantic alignment, making them effective auxiliary inputs for guiding subsequent image generation tasks. Figure 3.4 shows some results generated by Informative Drawings model.

Figure 3.4: Results generated by the Informative Drawings model. The inputs for these four images are the RGB outputs shown in Figure 3.3, and the generation is performed using the Informative Drawings model with anime-style weights.

Chapter 4

System Design

This study proposes a real-time guidance sketch generation system designed to assist users in improving drawing efficiency and creative quality. The system framework is built upon a diffusion model, taking the user's hand-drawn sketch and prompt as input conditions to generate RGB images that align with the user's creative intent. A Sketch Generator module is then used to extract a rough sketch from the generated image, which is further refined by the Denoising Optimizer module to produce a clearer and more structured guidance sketch. To support diverse artistic needs, the framework incorporates a Style-LoRA module, enabling the generation of guidance sketches in various drawing styles. Section 4.1 introduces the user interface design of the drawing assistance system. Section 4.2 details the processes of RGB image generation, sketch extraction, and sketch optimization. Figure 4.2 illustrates the overall architecture of the proposed system.

4.1 Interface Design

To enhance usability and creative flexibility for beginners in anime illustration, we have designed and implemented a feature-rich, user-friendly drawing assistance interface. This interface not only provides essential drawing tools but also integrates multiple guidance mechanisms. Additionally, it allows users to flexibly switch between different drawing styles, thereby improving both accessibility and expressive freedom. The interface layout is shown in Figure 4.1.

The proposed interface supports two primary input modes: freehand sketching using a brush tool and prompt input. When users find it difficult to clearly express their ideas using only a sketch, the prompt serves as a natural language input to guide the image generation process in alignment

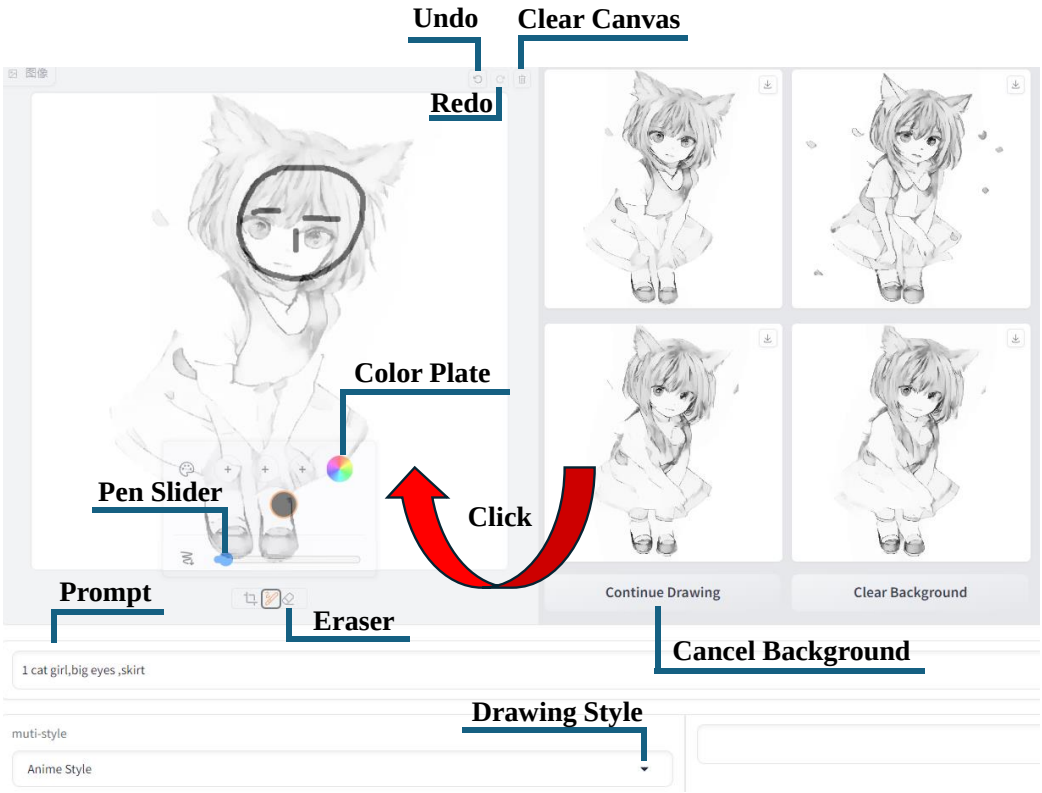


Figure 4.1: The interface provides a set of essential drawing tools, allowing users to flexibly manage background guidance sketches using the "Continue Drawing" and "Clear Background" buttons, and to express their intent through customizable prompts and selectable drawing styles.

with the user's creative intent. Furthermore, the system offers a set of predefined drawing styles, such as anime and realistic. Users can quickly select a preferred style, and the system will automatically adjust the generation strategy and model parameters to suit the chosen task.

Unlike traditional drawing assistance systems, which often fix the guidance sketch underneath the canvas and may constrain user creativity by encouraging tracing, our interface displays the guidance sketch on the right side of the canvas by default. Users can freely refer to the sketch without drawing directly over it. If tracing is desired, users can click on any guidance image to fix it as the canvas background, during which the system enters a paused state to facilitate accurate drawing. The background image can be removed via the "Clear Background" button, and generation remains paused until the user explicitly clicks "Continue Drawing" to resume. This flexible mechanism accommodates both reference-based and freeform drawing

modes, minimizing interference and promoting user autonomy.

The interface also includes basic features such as undo, redo, canvas clearing, eraser, color palette, and brush size adjustment, fulfilling a wide range of user drawing needs. To ensure smooth interaction, the system updates the guidance sketch only after each completed brush stroke, thereby avoiding frequent interruptions during the drawing process. To further enhance personalization and stylistic diversity, the system incorporates a drawing style selection list. Users can choose from multiple pretrained artistic styles (e.g., "anime" or "realistic"), and the system dynamically loads the corresponding Style-LoRA weights to generate guidance sketches in the selected style. This functionality significantly expands the creative potential of the system, supporting a wide variety of drawing scenarios. Overall, the proposed interface offers a comprehensive suite of drawing tools and controls, enabling flexible, precise, and real-time sketch guidance that is particularly well suited for beginners.

4.2 Image Generation

The proposed drawing assistance system first generates an RGB image that aligns with the user-selected drawing style, then extracts and refines a sketch from the image to produce a high-quality guidance sketch. Section 4.2.1 introduces the process of RGB image generation. Section 4.2.3 explains how the stylization module enables the production of images in various artistic styles. Section 4.2.4 provides a detailed description of the Sketch Generator module, which extracts a rough sketch from the RGB image, and the Denoising Optimizer module, which further refines the sketch to deliver a structurally clear and stylistically consistent guidance sketch to the user.

4.2.1 Multi-Sketch Generation

To enable the generation of diverse guidance sketches, this study proposes an image generation strategy that integrates multiple inference pipelines with conditional control mechanisms. By leveraging the parallel processing capabilities of GPUs, the system is designed to simultaneously execute four independent diffusion inference pipelines. While sharing identical model parameters and style configurations, each pipeline is configured with a unique combination of denoising timesteps. The hand-drawn sketch is first encoded into a latent representation and then denoised within each pipeline according to the respective timestep settings, resulting in multiple RGB images that are stylistically consistent but structurally varied. By exploiting the diver-

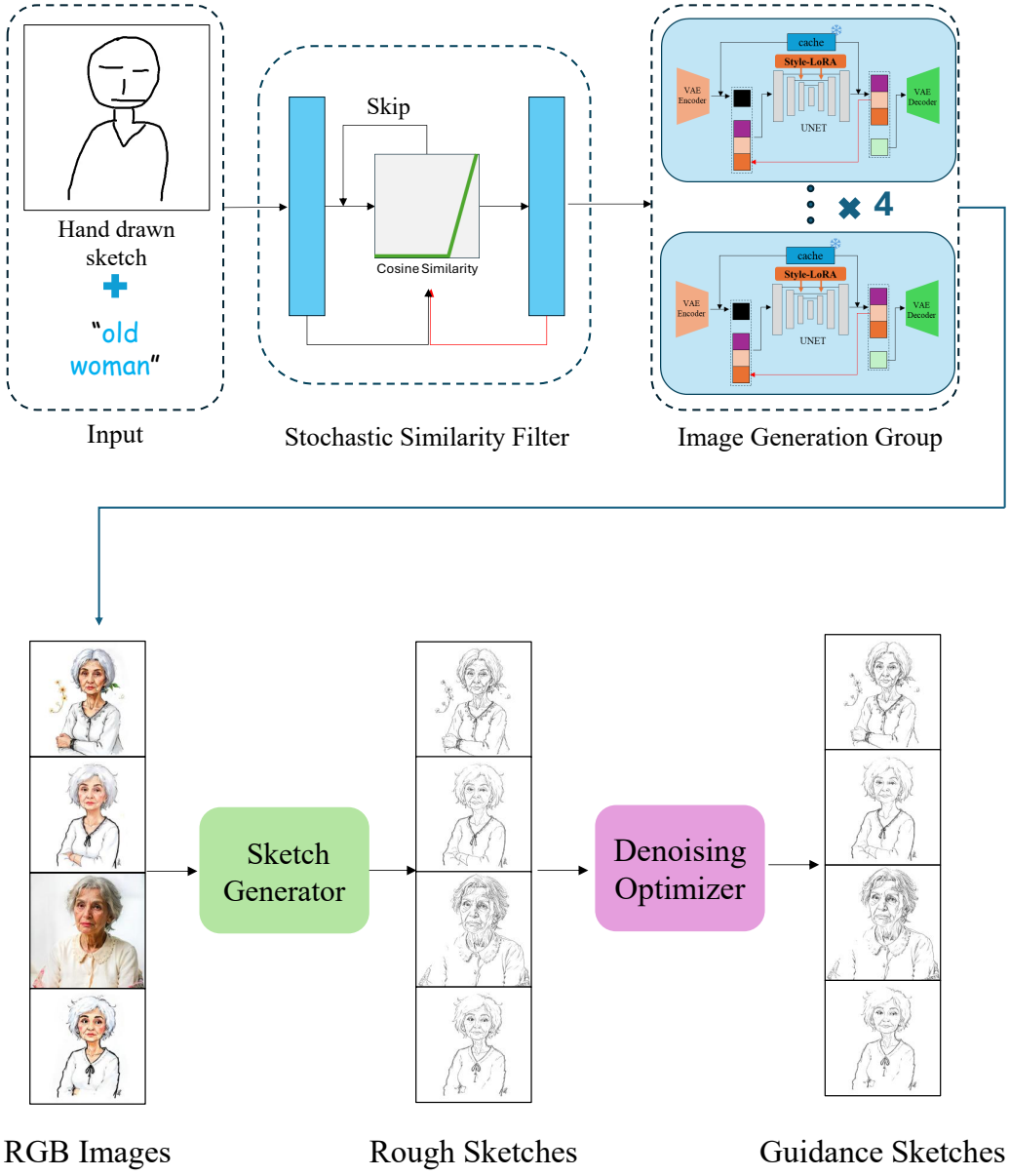


Figure 4.2: The framework of proposed drawing assistance system. The system takes a hand-drawn sketch and prompt as input. A similarity filter may skip generation if inputs are redundant. Four diffusion pipelines with different denoising steps produce diverse RGB images, which are converted to rough sketches and refined into clear guidance sketches for user assistance.

sity introduced by different timestep settings, the system effectively expands the range of generated outputs, thereby providing users with more diverse and personalized drawing references, enhancing both creative flexibility and interactive experience.

Beginners often struggle to accurately express their creative intent through hand-drawn sketches alone. Consequently, systems that rely solely on sketches for image generation may produce guidance sketches that deviate significantly from user expectations. To address this issue, this study incorporates prompts as auxiliary conditions during image generation. By introducing semantic guidance, the system enhances its capacity to capture user intent, thereby improving the accuracy and relevance of the generated outputs.

Specifically, the prompt is first encoded into a high-dimensional semantic embedding via a text encoder, and subsequently injected as conditional information into the cross-attention module of the U-Net, thereby reinforcing the model's sensitivity to user semantic intent. In the cross-attention mechanism, the association between the image-encoded query $\mathbf{Q}_{\text{image}}$ and the prompt-encoded key and value representations $\mathbf{K}_{\text{prompt}}$ and $\mathbf{V}_{\text{prompt}}$ is computed as follows:

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

where d_k denotes the dimensionality of the key vectors. This mechanism enables the model to dynamically focus on regions of the image that are semantically aligned with the user's prompt, thereby mitigating ambiguity caused by imprecise sketches.

To further modulate the model's reliance on the prompt, the system introduces a Classifier-Free Guidance (CFG) mechanism. By blending the noise predictions from the prompt-conditioned and non-conditioned branches, the model achieves flexible control over guidance strength. The blending formula is defined as:

$$\hat{\varepsilon} = \varepsilon_{\text{no prompt}} + w \cdot (\varepsilon_{\text{prompt}} - \varepsilon_{\text{no prompt}})$$

Here, $\varepsilon_{\text{prompt}}$ represents the noise prediction from the branch conditioned on the prompt, $\varepsilon_{\text{no prompt}}$ denotes the prediction without prompt conditioning, and w is the guidance scale coefficient. By adjusting the value of w, users can flexibly control the influence of the prompt during image generation, thereby supporting a spectrum of use cases from weak to strong guidance, suitable for both novice and advanced users at different stages of the creative process. The RGB image generated by the proposed system is shown in Figure 4.3.

Figure 4.3: RGB images generated by the proposed drawing assistance system without applying any Style-LoRA. All images were generated using prompts only, with Stable Diffusion 1.5 model weights and four denoising steps, at a resolution of 512×512. The prompts for the first row are "a girl", "a boy", and "bear", respectively. The prompts for the second row are "flower", "river", and "garden".

4.2.2 Structure-Preserving Denoising with Sketch Inputs

User-provided hand-drawn sketches typically embody clear creative intent. Therefore, it is essential for the guidance sketches generated by the drawing assistance system to maintain structural consistency with the input sketches. To effectively preserve structural information, the proposed system incorporates residual and skip connection mechanisms during the denoising process, thereby enhancing the model's ability to retain the structure of the sketches. Specifically, the hand-drawn sketch is first encoded into a latent representation using a Variational Autoencoder (VAE), and Gaussian noise is added at a specified timestep to simulate an intermediate state in the diffusion process. The resulting noisy latent is then fed into the U-Net for denoising reconstruction.

The U-Net extracts multi-scale features from the hand drawn sketch

through a series of convolutional layers. Residual connections are introduced within each convolutional block, such that each layer's output satisfies:

$$h(x) = f(x) + x$$

where x denotes the hand-drawn sketch feature, and $f(x)$ represents the output of the convolutional transformation. This design not only enhances the stability of feature propagation but also effectively preserves local structural information embedded in the input sketch.

In addition, the system incorporates skip connections that directly transfer intermediate features from the downsampling path to their corresponding layers in the upsampling path. This mechanism facilitates the integration of low-level detail with high-level semantic representations, thereby ensuring that the final generated image remains structurally aligned with the user's input sketch.

4.2.3 Drawing Style

In drawing assistance systems, the accuracy of visual style plays a crucial role in providing meaningful guidance. For example, in the case of the anime-style, models must exhibit strong stylistic fidelity and sensitivity to visual details. However, training large diffusion models from scratch for specific styles demands extensive data and computational resources. To address this challenge, we adopt LoRA to fine-tune the weights of a pre-trained model, enabling efficient and parameter-saving multi-style generation.

As a representative case, we curated a dataset of 40 high-quality anime-style images to facilitate precise and diverse style learning, including 17 images from the Pixiv platform [48] and 23 from PixAi [49]. The dataset includes 26 images of human characters with various genders, poses, and facial expressions, and 14 images of animal characters, including anthropomorphic figures such as kemonomimi (animal-eared) characters. The images span everyday, fantasy, and school uniform themes and exhibit classic anime features such as large eyes, strong outlines, and expressive emotions. During training, the parameters of the backbone network remain frozen, and model adaptation is achieved solely by updating the weights of the LoRA modules. Upon completion of training, the learned LoRA weights are integrated into the attention layers of the model to enable style-specific guidance. This approach allows the proposed system to flexibly support multiple artistic styles without retraining the entire backbone, thereby significantly enhancing its scalability and adaptability to diverse drawing assistance scenarios. Figure

4.4 illustrates example outputs produced with our anime Style-LoRA, showing strong stylistic alignment and fine detail generation.

Figure 4.4: RGB images generated by the proposed system with the anime Style-LoRA applied.

4.2.4 Sketch Generation and Optimization

The proposed system adopts the Informative Drawings model based on a symmetrical network architecture. The model takes a RGB image as input and outputs a single-channel rough sketch, using a Sigmoid activation function to normalize the pixel values to the range of 0 to 1.

In the encoding stage, the model first applies a 7×7 convolution to perform initial feature extraction, followed by downsampling the input using a stride-2 convolution layer to expand the receptive field and capture

multi-scale feature representations. The feature transformation process incorporates multiple residual blocks. Each residual block consists of two 3×3 convolution layers, normalization layers, and the ReLU activation function [50] to enhance structural representation and effectively preserve geometric and edge information in the image. During the decoding stage, two transposed convolution layers are employed to perform upsampling and progressively restore the spatial resolution of the image, ultimately producing the output sketch.

Since the initially generated rough sketch often contains noise and redundant details, a denoising optimizer is introduced to further enhance the usability and clarity of the sketch. This optimizer is based on a recursive filtering mechanism, which preserves major structural edges while effectively suppressing high-frequency noise and unnecessary textures, thereby improving the overall continuity and visual quality of line drawings. The denoising optimizer operates by performing one-dimensional filtering along both horizontal and vertical directions, resulting in two-dimensional edge-preserving smoothing. Compared to traditional Gaussian filters, the proposed optimizer offers superior edge retention while reducing minor noise, producing cleaner and more coherent contour lines. This, in turn, enhances the visual clarity and structural readability of the sketch, providing more reliable guidance for downstream user drawing. Figure 4.5 shows the smoothed image and the corresponding line drawing and RGB image.

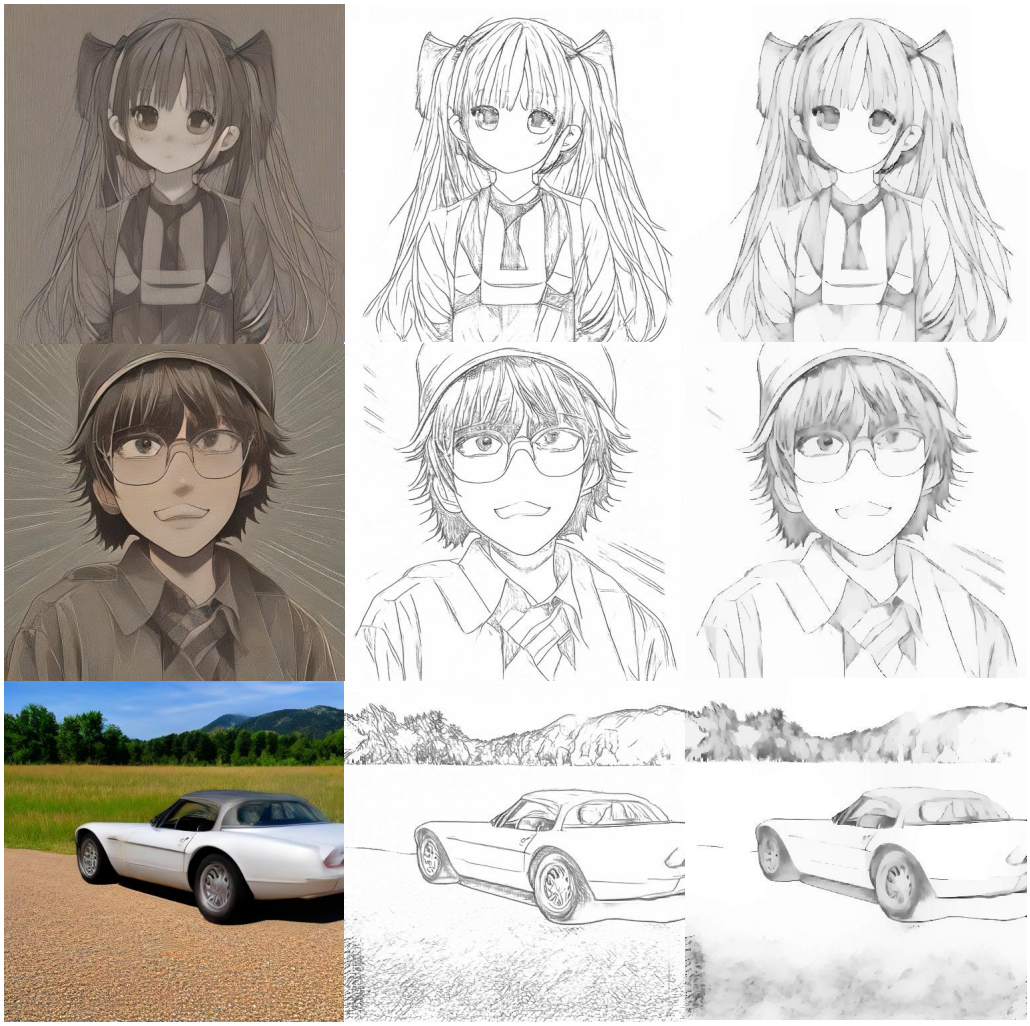


Figure 4.5: The first column shows RGB images generated by the proposed drawing assistance system. The first two rows adopt the anime Style-LoRA, while the third row uses the realistic style. The second column presents the corresponding rough sketches extracted by the Sketch Generator. The third column shows the refined guidance sketches after processing with the denoising optimizer.

Chapter 5

User Study

To evaluate the usability of the proposed drawing assistance system, a user study was designed and conducted, focusing on participants' subjective experiences. The study included post-usage ratings and preference assessments provided by users. Through the collection and analysis of experimental data, this research systematically and quantitatively assessed the effectiveness of the proposed system in supporting drawing tasks, as well as the usability and user acceptance of its interface design. This chapter provides a comprehensive overview of the design and implementation of the user study. Section 5.1 details the composition of the participants, the data selection criteria, and the experimental platform; Section 5.2 presents the experimental design, including interface configuration, questionnaire structure, drawing subject selection, and procedural workflow.

5.1 User Study Background

This section introduces the participants involved in the experiment, the data screening procedures implemented to ensure the fairness and validity of the results, and the experimental equipment used. Section 5.1.1 presents the detailed background and composition of the participants; Section 5.1.2 outlines the experimental environment and hardware configuration used throughout the study.

5.1.1 Participants

A total of eight participants were recruited for this user study, including one female and seven males. To comprehensively evaluate the applicability and effectiveness of the proposed drawing assistance system across diverse

user groups, the participant pool included both users with professional art backgrounds and those without prior drawing experience. Specifically, two participants had received systematic training in fine arts for several years and held bachelor's degrees in art-related disciplines; these individuals were classified as expert users. The remaining six participants, who had not undergone any formal training in drawing, were categorized as non-expert users. All participants were current master's students with basic proficiency in digital tools and independent communication skills, enabling them to accurately report their user experiences and subjective impressions.

5.1.2 Experimental Setup

The hardware platform used for this experiment consisted of a desktop computer running Windows 11 Home Edition, equipped with an AMD Ryzen 5 4500 six-core processor and an NVIDIA GeForce RTX 4060 GPU. This local machine was connected via a network to a high-performance remote server operating on Ubuntu, configured with an NVIDIA GeForce RTX 4090 GPU and an Intel Core i7-13700KF processor.

All system modules, including core computational components such as image generation and sketch extraction, were deployed and executed on the remote server. During the experiment, participants interacted with the system through the local terminal. The user interface of the drawing assistance system was built on the Gradio platform and executed locally to receive user input and display the generated outputs. After completing computational tasks on the server, the system transmitted the results back to the local machine, where they were displayed in real-time via the Gradio interface, thereby enabling efficient human–computer interaction.

All drawing input was conducted using a WACOM graphics tablet, with the canvas resolution uniformly set to 512×512 pixels to ensure consistency across experimental conditions.

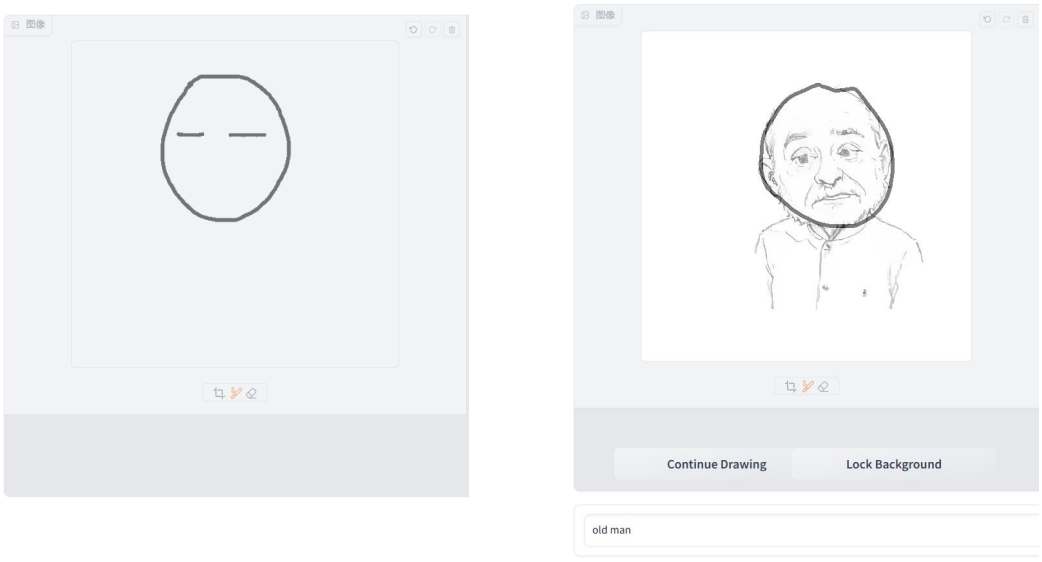
5.2 User Study Design

This section systematically presents the design of the user study conducted for the proposed drawing assistance system. Section 5.2.1 introduces the three drawing interfaces used in the experiment and their key interaction differences. Section 5.2.2 details the questionnaire design, including the SUS used to evaluate system usability and preference-based questions to assess user subjective choices. Section 5.2.3 explains the rationale behind the selection of drawing subjects, with an emphasis on the decision to adopt realistic-

style illustrations. Section 5.2.4 describes the overall experimental procedure in detail.

5.2.1 Interface Design

The user study employed a comparative experimental design to evaluate how different drawing assistance interfaces influence user drawing behavior and experience. Three distinct interfaces were developed for this purpose: the Baseline Interface, the Shadow Guidance Interface, and the Proposed Interface. A systematic user study was conducted to assess the impact of each interface on participants' drawing performance. the Baseline Interface and the Shadow Guidance Interface are shown in Figure 5.1.



(a) (b)

Figure 5.1: Drawing interfaces used in our study. (a) Baseline interface: a canvas without user guidance; (b) Shadow guidance interface: interface with one guidance sketch places under the canvas.

The Baseline Interface served as a control condition, where participants drew freely on a blank canvas without any assistance. This setup was intended to assess users' inherent drawing abilities without system intervention.

Both the Shadow Guidance Interface and the Proposed Interface were built upon the same drawing assistance system developed in this study. They utilized identical guidance sketch generation mechanisms to ensure that variations in drawing outcomes were attributable to differences in interface design

rather than sketch quality. This allowed the experiment to specifically focus on how different interface layouts affect user behavior and their ability to express creative intent—namely, which interface best supports the production of drawings aligned with users' envisioned outcomes, rather than simply evaluating final drawing quality.

In the Shadow Guidance Interface, the system dynamically generates guidance sketches based on the user's drawing input and textual prompts. These sketches are displayed directly in the background layer of the canvas, allowing users to trace over them as they draw.

The Proposed Interface offers the same sketch generation capability, but differs in layout: it separates the canvas and guidance display areas. Specifically, the interface presents four candidate sketches on the right-hand side of the canvas. Users can select any of these as reference images and overlay one as the canvas background to assist their drawing. At any time during the drawing process, users may remove the guidance sketch to return to a blank canvas or switch to a different reference sketch. This interactive mechanism allows users to flexibly add, remove, or swap guidance sketches throughout their creative process. Alternatively, users may choose to use the guidance sketches purely as visual references or disregard them entirely and draw independently.

To minimize distractions caused by frequent sketch updates, both the Shadow Guidance and Proposed Interfaces include a "pause" function. This feature allows users to freeze a selected guidance sketch once they are satisfied with it, keeping it fixed in the background while they continue drawing. This mechanism helps maintain visual consistency, enhances creative focus, and improves the overall continuity and controllability of the drawing experience.

5.2.2 Questionnaire Design

System Usability Scale

The usability of a drawing assistance system has a direct and significant impact on the user experience. To systematically evaluate the usability of the proposed system in this study, we employed the System Usability Scale (SUS) for quantitative analysis. The SUS questionnaire consists of ten standardized questions, including five positively worded items and five negatively worded items, each designed to assess different aspects of users' subjective experiences with the proposed system. The positively worded items are primarily intended to measure users' favorable experiences. For example, the question "I think that I would like to use this system frequently." is designed to assess users' willingness to continue using the system; "I thought the sys-

tem was easy to use." evaluates the system's ease of use; and "I found the various functions in this system were well integrated." focuses on the integration and coordination among different system modules, reflecting whether the system meets users' diverse needs during the creative process. The negatively worded items are designed to identify potential usability issues. For instance, "I found the system very cumbersome to use." explores whether users perceive the proposed system as complicated during operation, while "I needed to learn a lot of things before I could get going with this system." evaluates the learning cost and the difficulty of getting started. These questions collectively provide a comprehensive assessment of system usability and offer a reliable data for evaluating the proposed drawing assistance system in this study. The specific content of the SUS form is presented in Table 6.1.

User Preference Questions

To further investigate the impact of different interface interaction designs on users' drawing behaviors and experiences, the questionnaire included a set of user preference questions. Specifically, this study aimed to explore which type of interaction design better enables users to produce drawings that align with their original creative intentions, rather than merely copying images that do not reflect their true goals. For example, questions such as "Which experiment's results best matched your initial expectations?" were designed to collect participants' subjective preferences and experiences with the three drawing interfaces. The complete list of preference questions is presented in Table 6.2.

These questions were presented in the form of single-answer multiple-choice questions, and participants were asked to select the interface that best matched their experience or satisfaction. By analyzing the distribution of participants' responses, this study aims to determine which interface design more effectively supports users in achieving their creative objectives, thereby providing valuable insights for future improvements in the design of drawing assistance system interfaces.

5.2.3 Drawing Subject Selection

The user study primarily focused on realistic-style drawing tasks. Participants were asked to select three drawing topics from a predefined list of five. The realistic style was chosen as the primary drawing theme because it requires minimal prior domain knowledge, allowing for a more standardized experimental baseline among participants.

Other styles, such as anime, were excluded from the formal analysis due to their high dependence on stylistic conventions and cultural familiarity. For users unfamiliar with these conventions, accurately expressing their creative intentions could be challenging, thus introducing potential bias into the experimental results. While some participants were invited to experiment with anime-style drawing during the study, the corresponding results were not included in the formal data analysis.

5.2.4 Experimental Procedure

In the experimental procedure, each participant was required to complete three drawing tasks using the three different interface conditions. All participants were instructed to begin with the Baseline Interface for the first drawing task. This arrangement served two purposes: first, to evaluate each participant's baseline drawing ability without system assistance as a reference; and second, to prevent any prior exposure to system-generated guidance from influencing their initial creative thinking, thus ensuring the fairness of the experimental conditions.

After completing the task using the Baseline Interface, participants proceeded to complete the remaining two tasks using the Shadow Guidance Interface and the Proposed Interface, with the order of these two conditions randomized among participants to mitigate potential order effects.

Before the formal experiment, all participants were given time to familiarize themselves with the features and interaction mechanisms of all three interfaces to ensure that they could complete the tasks independently. Upon completing all drawing tasks, participants were asked to complete a questionnaire evaluating the overall usability of the system and the effectiveness of each interface in supporting their drawing process. The questionnaire included the SUS and a set of user preference questions, aimed at comprehensively assessing system performance, user experience, and the impact of interface design on drawing outcomes.

5.3 Drawing Quality Evaluation Experiment

To systematically assess the effectiveness of the proposed drawing assistance system in improving users' drawing quality, this chapter presents an objective evaluation of the artworks produced by participants under different interface conditions. By involving evaluators from both expert and non-expert backgrounds and employing a set of predefined multi-dimensional evaluation criteria, we comprehensively analyzed the drawing outcomes across interfaces to

validate the practical utility and guidance capability of the proposed system. Section 5.3.1 introduces the composition and background of the evaluators involved in the assessment task, while Section 5.3.2 details the evaluation metrics used to assess drawing quality.

5.3.1 Evaluator Background

A total of 13 evaluators were invited to assess the quality of participants' drawings created under the three interface conditions. This group included 10 non-expert evaluators and 3 expert evaluators. None of the evaluators had participated in the user study, ensuring the objectivity and impartiality of the evaluation process.

The 10 non-expert evaluators lacked formal training in drawing and represented the general audience perspective, providing intuitive assessments of drawing appeal and clarity. The 3 expert evaluators had backgrounds in the arts, including university instructors specializing in art education and professionals with degrees from art institutions currently working in creative industries. These experts offered more informed evaluations based on structural coherence, proportion, and expressive quality. This diverse composition of the evaluation panel was intended to integrate both lay and expert perspectives, enabling a comprehensive and balanced analysis of drawing quality across multiple dimensions.

5.3.2 Evaluation Criteria

To evaluate the quality of users' drawings across different interface conditions, three core evaluation metrics were established: Overall Shape, Local Proportion, and Line Quality. Each metric was assessed using a five-point Likert scale, with 5 indicating excellent performance and 1 indicating the lowest level. Overall Shape assesses whether the drawing has a clear composition and coherent outline, reflecting the participant's ability to grasp global form. Local Proportion focuses on the accuracy and harmony of proportions within facial features or body parts, indicating the participant's control over detailed structure. Line Quality evaluates the smoothness, neatness, and expressive strength of the lines, reflecting the user's technical fluency and visual clarity during the drawing process.

This evaluation design not only enables a direct comparison of the guidance effects across different interfaces but also facilitates the analysis of how interface interaction design influences users' ability to express their drawing intentions and maintain creative autonomy. Furthermore, it provides insights into how the proposed system supports novice versus expert users differently,

offering empirical evidence for future improvements in interface design and model development.

Chapter 6

Results

To evaluate the proposed drawing assistance system from both subjective and objective perspectives, this study analyzed questionnaire data collected during the experiment and conducted an assessment of participants' completed drawings. Section 6.1 presents the results of the system usability evaluation. Section 6.2 reports on how different interface designs influenced participants' drawing performance. Section 6.3 presents the drawing quality scores. Finally, Section 6.4 provides a qualitative analysis based on participants' behavior and feedback during the drawing process, offering further insights into human–computer interaction experiences and creative autonomy across the three interfaces.

6.1 System Usability Analysis

To ensure reliable analysis, we first screened the collected user study data. If a participant gave the same response to all questionnaire items, their data were considered invalid due to lack of engagement. One participant gave the same response to all questionnaire items and was excluded. The final analysis was based on valid responses from the remaining seven participants. Table 6.1 presents the SUS scores obtained during the user study. The scores are rounded to two decimal places. According to established guidelines, an average SUS score of 85 or higher is generally interpreted as indicating excellent usability. In this study, the proposed system achieved an average score of 84.67, which is close to the threshold for excellence and demonstrates strong overall usability. In particular, Question 5, which asked whether "the functions of the system were well integrated," received a score of 4.57. This indicates that users perceived the system as having a high degree of functional integration. Furthermore, Question 1, which asked whether "the user

would like to use this system frequently," received a score of 4.43, suggesting that participants found the system intuitive and easy to adopt. For the negatively worded items such as Question 8, which stated "I found the system very cumbersome to use," the score was as low as 1.43. This further supports the conclusion that the system was perceived as user-friendly and efficient. These results demonstrate that participants found the proposed drawing assistance system to be both user-friendly and efficient. The SUS results confirm that users were able to operate the system easily and use it effectively to support their drawing tasks. Figure 6.1 shows the drawing results of different interfaces used in the user experiment.

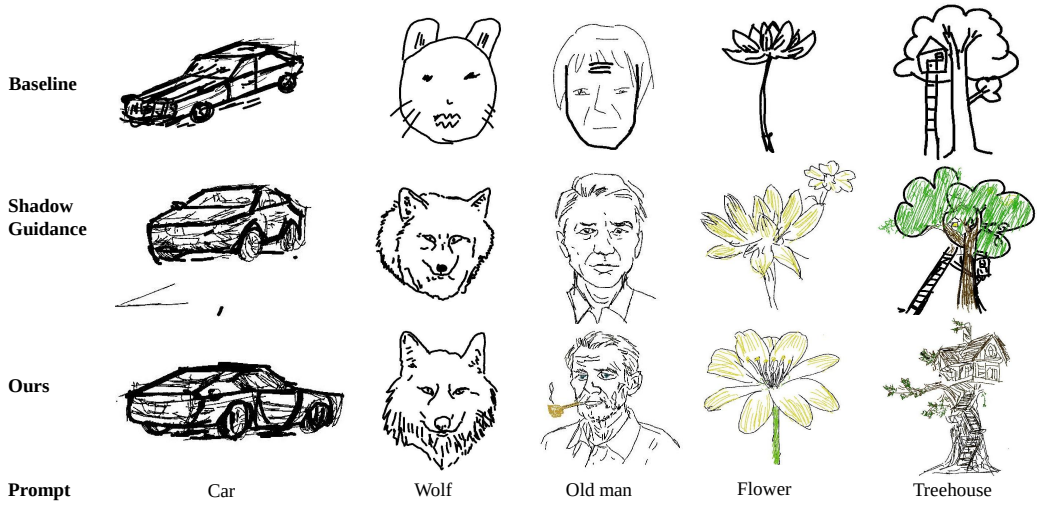


Figure 6.1: The figure presents the results of the user study. The first column illustrates samples created by a professional user, while the remaining columns display drawings produced by novice users. All participants generated their illustrations using the "realistic style" option provided by the system.

6.2 User Preference Analysis

In order to verify which interface better enables users to produce drawings that align with their intended expectations, this study incorporated user preference questions into the questionnaire and conducted a statistical analysis of the participants' responses. As shown in Table 6.2, the proposed interface outperformed the other two interfaces in terms of user satisfaction and drawing assistance effectiveness. For the question "Which interface best matched

Table 6.1: Results of the SUS questionnaire. ⇑ indicates that higher scores are better; ⇓ for the other case.

#	Questions	Mean	SD
1	I would like to use this system frequently. ⇑	4.43	0.73
2	I found this system unnecessarily complex. ⇓	2.14	0.99
3	This system was easy to use. ⇑	4.43	0.73
4	I would need the support of a technical person to be able to use this system. ⇓	1.71	1.16
5	I found the various functions in this system were well integrated. ⇑	4.57	0.49
6	I thought there was too much inconsistency in this system. ⇓	1.57	0.49
7	I would imagine that most people would learn to use this system very quickly. ⇑	4.43	0.49
8	I found the system very cumbersome to use. ⇓	1.43	0.73
9	I felt very confident in using this system. ⇑	4.43	0.49
10	I needed to learn a lot of things before I could get going with this system. ⇓	1.57	1.05

your initial expectations?", 85.7% of participants selected the proposed interface, 14.3% selected the Shadow Interface, and no participants selected the Baseline Interface. This indicates that the interaction mode of the proposed interface aligns more closely with users' creative thinking and helps guide them to produce drawings that meet their initial expectations. For the question "Which interface was the most helpful for your drawing?", 71.4% of participants considered the proposed interface to be the most effective, followed by 28.6% who chose the Shadow Interface, while the Baseline Interface received no votes. These results further support the practicality and effectiveness of the proposed interface.

6.3 Drawing Results Evaluation

To objectively assess the effectiveness of different interfaces in enhancing drawing quality, this study invited 13 evaluators (including 10 non-experts and 3 experts) to rate the drawings produced by participants. The evaluation criteria included Overall Shape, Local Proportion, and Line Quality. The detailed scoring results are presented in Table 6.3. The scoring data indicate

Table 6.2: The result of user preference experiment.

Questions	Baseline	Shadow	Ours
Which experiment's results best matched your initial expectations?	0	14.3%	**85.7%**
Which is your desired outcome in the three groups?	0	28.6%	**71.4%**
Which one do you think is better for you to draw?	0	14.3%	**85.7%**

that the proposed interface outperformed both the Baseline Interface and the Shadow Interface across all evaluation metrics. For instance, in the Overall Shape dimension based on non-expert evaluations, the proposed interface achieved a score of 4.40, which was significantly higher than the Baseline's 2.20 and the Shadow Interface's 3.07. This demonstrates that the proposed interface notably improved users' ability to grasp structural form. Expert evaluations followed a similar trend, further validating the effectiveness of the proposed system.

In summary, the quantitative analysis clearly demonstrates that the drawing assistance system proposed in this study excels in improving both user experience and drawing quality. The proposed interface not only enhanced the usability of the system but also effectively supported users in expressing their creative intentions, providing substantial assistance to both novice and experienced users.

6.4 Qualitative Analysis Results

When using the traditional non-assisted interface (Baseline Interface), novice users exhibited notable uncertainty and frequent revisions during the drawing process. They often struggled with capturing accurate proportions and overall composition, resulting in immature strokes and prolonged drawing times. Despite their efforts, the final outputs were frequently structurally loose and lacked visual depth.

In comparison, the Shadow Guidance Interface, which provides a fixed sketch reference in the background layer, helped beginners establish basic shape contours. However, due to the sketch being forcibly fixed to the canvas

background, some users reported that the provided guidance sketch interfered with their drawing process, leading them to rely more on passive tracing rather than active creation.

The proposed drawing assistance interface significantly improved users' sense of agency and flexibility. It allows users to freely select reference sketches from a panel of candidates on the right side and decide whether to fix a chosen sketch as the canvas background. Users could reference the parts that aligned with their intended outcome and ultimately complete a drawing that reflected their own vision. Most users responded that this approach "better aligned with their original drawing intent."

Overall, users experienced a more fluid and autonomous creative process when using the proposed drawing assistance interface. The final drawings were also more consistent with their creative intentions. These qualitative observations are consistent with the questionnaire data, further validating the practical advantages of the system in terms of human-computer interaction experience and drawing efficiency. Figure 6.2 showcases the drawing results produced under the "anime-style" setting across different interface conditions.

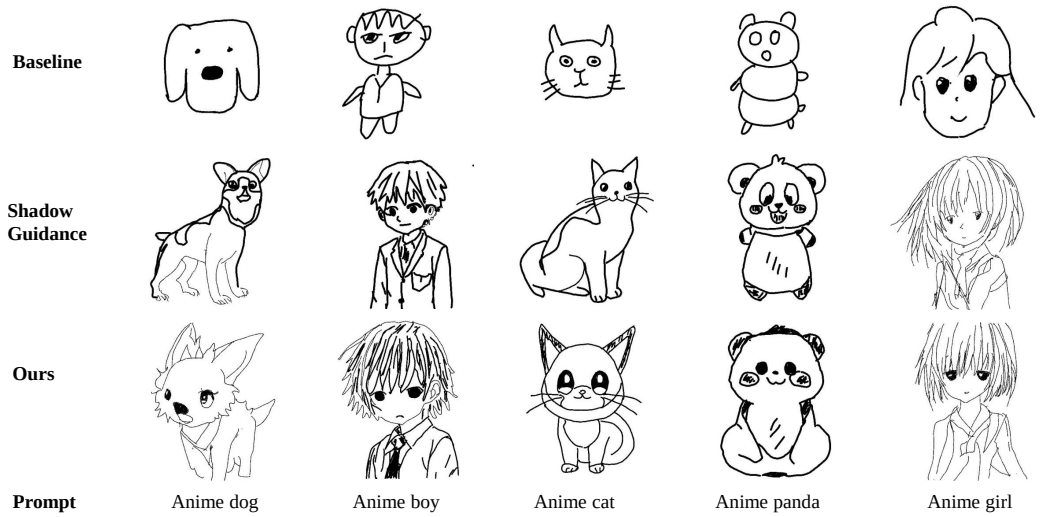


Figure 6.2: Anime-style drawing assistance results. These examples serve to demonstrate the system's support for diverse artistic styles and are not included in the experimental evaluation. In the figure, the first row displays drawings generated using the Baseline Interface, the second row shows results obtained with the Shadow Guidance Interface, and the third row presents outputs produced by the proposed interface.

Table 6.3: The drawing result scores for overall shape, local proportion, and line quality across the baseline interface, shadow guidance interface, and our interface for non-experts and experts.

Methods	**Group**	**Overall Shape**	**Local Proportion**	**Line Quality**
Baseline	Non-expert	2.20	2.18	2.58
	Expert	2.40	2.40	2.70
Shadow	Non-expert	3.07	3.11	3.40
	Expert	3.30	3.50	3.80
Ours	Non-expert	**4.40**	**4.36**	**4.29**
	Expert	**4.10**	**4.20**	**4.40**

Chapter 7

Conclusion

This chapter summarizes the study and highlights the key experimental results of the proposed system. Section 7.1 introduces the design and implementation of the real-time drawing assistance system based on diffusion models, and presents the evaluation results from usability testing, user preference analysis, and drawing quality assessments, demonstrating that the system significantly improves drawing quality and facilitates creative expression, particularly benefiting beginners; finally, Section 7.2 discusses the current limitations of the system, including its limited ability to preserve fine-grained local details and the lack of support for incremental local updates, and proposes potential directions for future enhancement.

7.1 Summary

This study proposes a drawing assistance system based on diffusion model, designed to provide users with real-time, high-quality guidance sketches across multiple artistic styles. The system takes hand-drawn sketches and textual prompts as inputs and employs the StreamDiffusion inference framework in conjunction with Style-LoRA modules to generate multi-style guidance sketches in real time. A rough sketch is first extracted through a sketch generator and then refined by a denoising optimizer, which removes noise and enhances edge details, ultimately producing structured and detailed guidance sketches. Compared to conventional drawing assistance methods, the proposed system offers significant advantages in both interactivity and style adaptability. It can generate four guidance sketches in real time, allowing users to select a preferred sketch and set it as the background for tracing, thereby providing references while preserving creative freedom. Furthermore, the system supports a variety of drawing styles, enabling users to switch flex-

ibly according to their creative needs. Results of this study demonstrate that the proposed system effectively enhances drawing efficiency and quality, particularly benefiting beginners. Users are able to better express their intended ideas and produce drawings that more closely align with their expectations.

To evaluate the usability and authoring flexibility of the proposed system, we conducted a user study that combined a quantitative usability evaluation of the interface with an analysis of user preferences.The experiment compared three distinct drawing interfaces: a baseline interface without any assistance, a shadow guidance interface with the guidance sketch fixed on the background layer, and our proposed interactive interface offering multiple selectable guidance sketches. In terms of usability assessment, the system achieved an average score of 84.67 on the System Usability Scale (SUS), indicating excellent performance in user experience and functional integration, and demonstrating strong overall usability. Additionally, the results of the user preference survey showed that the majority of participants favored our proposed interface, citing its superior ability to express drawing intentions and support creative freedom. In the drawing quality evaluation, sketches produced using our system outperformed those from the other interfaces in terms of overall shape, local proportions, and line clarity. The system notably improved drawing outcomes, particularly for novice users, enabling them to create structurally sound and detailed sketches within a shorter time frame.

In summary, the experimental results provide strong evidence for the effectiveness and practical value of the proposed drawing assistance system in enhancing drawing efficiency, improving sketch quality, and supporting user creativity.

7.2 Limitations

Although the proposed drawing assistance system performs well in providing overall structural guidance, it still exhibits certain limitations in detailed expression and local control. Specifically, when users include local features such as decorative elements or unique structures in their hand-drawn sketches, the system may fail to accurately preserve and reflect the intended creative details. As illustrated in Figure 7.1, when a user adds elements like horns to a character's head, the generated guidance sketch may omit these features, resulting in a mismatch between the user's input and the generated output. This indicates that the current global generation mechanism has limitations in handling fine-grained details, potentially leading to semantic inconsistencies between the hand-drawn sketch and the corresponding guidance sketch.

Moreover, the current system treats the entire image as a unified whole

and lacks support for local updates, making it difficult to maintain the stability of unmodified regions during partial edits. To address this issue, future work will explore the incorporation of local incremental update mechanisms, enabling regeneration of only the modified areas while preserving the consistency of the untouched regions. In addition, we plan to integrate conditionally controlled generation models such as ControlNet [51], which utilize structural information from the sketch as control signals to enhance the consistency and fidelity of the generated guidance sketches with respect to user input.

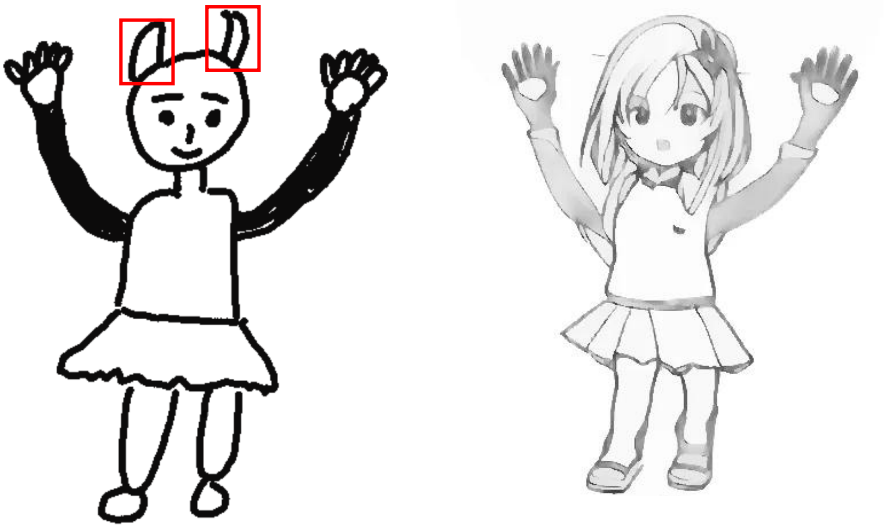


Figure 7.1: Example of a limitation case: The left image shows a hand-drawn sketch input by the user in the proposed system, with the red box highlighting the added detail of "horns". The right image presents the generated guidance sketch, in which the character's head fails to reproduce the horn feature. This indicates a limitation of the proposed system in preserving details in the generated guidance sketches.

Acknowledgement

During my two years as a master's student, I have been extremely fortunate to conduct my research under the guidance of Professor Xie. When I first arrived, I had little knowledge of research. Over the past two years, I have gradually acquired the ability to conduct academic research. The rigorous and honest academic atmosphere in Professor Xie's lab has not only shaped my research capabilities but also taught me valuable lessons about integrity and professionalism. What I had previously only heard as principles of life and scholarship, I have now experienced firsthand. This journey has left a deep impression on me and has served as a guiding light in both my academic and personal growth. I am sincerely grateful for Professor Xie's patient guidance and support throughout my studies.

I would also like to thank my fellow lab members. As someone with limited language proficiency when I first arrived, I often struggled in class. It was thanks to the help of my peers that I was able to overcome these difficulties and make steady progress. I am also grateful to my senior labmates for their advice and guidance in research, which helped me avoid many obstacles. My sincere appreciation goes to my advisor and all lab members.

Finally, I would like to thank my family for their unwavering support throughout this journey. Although we are separated by distance, their constant encouragement has been a source of strength and motivation. Without their support, I would not have achieved my current growth.

References

[1] Adobe Inc., "Adobe photoshop," retrieved January 29, 2025. [Online]. Available: https://www.adobe.com/jp/products/photoshop.html

[2] CELSYS, Inc., "Clip studio paint," retrieved January 29, 2025. [Online]. Available: https://www.clipstudio.net/ja/

[3] Savage Interactive, "Procreate," retrieved January 29, 2025. [Online]. Available: https://procreate.com/

[4] J. Xie, A. Hertzmann, W. Li, and H. Winnemöller, "Portraitsketch: Face sketching assistance for novices," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 407–417.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[6] I. J. 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.

[7] S.-H. Zhang, Y.-C. Guo, and Q.-W. Gu, "Sketch2model: View-aware 3d modeling from single free-hand sketches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6012–6021.

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

[9] C. Chen, X. Xie, Y. Zhang, T. Zhang, and H. Xie, "Interactive drawing guidance for anime illustrations with diffusion model," in *2025 Nicograph International (NICOInt)*. IEEE, 2025, pp. 1–8.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[11] MidJourney, Inc., "Midjourney," retrieved January 29, 2025. [Online]. Available: https://www.midjourney.com/

[12] OpenAI, "Dall·e," retrieved January 29, 2025. [Online]. Available: https://openai.com/dall-e

[13] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. Torr, and E. Shechtman, "Interactive sketch & fill: Multiclass sketch-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1171–1180.

[14] J. Choi, H. Cho, J. Song, and S. M. Yoon, "Sketchhelper: Real-time stroke guidance for freehand sketch retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2083–2092, 2019.

[15] J. Wang, K. Nakano, D. Chen, Z. Huang, T. Fukusato, K. Miyata, and H. Xie, "A study on cognitive effects of canvas size for augmenting drawing skill," in *2024 Nicograph International (NicoInt)*. IEEE, 2024, pp. 49–53.

[16] H. Kanayama, H. Xie, and K. Miyata, "Illustration drawing interface with image retrieval and adjustable grid guidance," in *2023 Nicograph International (NicoInt)*. IEEE, 2023, pp. 54–61.

[17] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

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

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

[21] 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.

[22] S. Luo, Y. Tan, S. Patil, D. Gu, P. Von Platen, A. Passos, L. Huang, J. Li, and H. Zhao, "Lcm-lora: A universal stable-diffusion acceleration module. arxiv 2023," *arXiv preprint arXiv:2311.05556*.

[23] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[24] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[25] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, 2017. [Online]. Available: http://arxiv.org/abs/1710.10196

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

[27] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational bayes," 2013.

[28] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

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

[30] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[31] 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.

[32] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," 2023.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 297–14 306.

[35] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in neural information processing systems*, vol. 35, pp. 5775–5787, 2022.

[36] Lu, Cheng and Zhou, Yuhao and Bao, Fan and Chen, Jianfei and Li, Chongxuan and Zhu, Jun, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, pp. 1–22, 2025.

[37] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa, "Photosketcher: interactive sketch-based image synthesis," *IEEE Computer Graphics and Applications*, vol. 31, no. 6, pp. 56–66, 2011.

[38] Y. J. Lee, C. L. Zitnick, and M. F. Cohen, "Shadowdraw: real-time user guidance for freehand drawing," *ACM Transactions on Graphics (ToG)*, vol. 30, no. 4, pp. 1–10, 2011.

[39] A. K. Bhunia, S. Koley, A. F. U. R. Khilji, A. Sain, P. N. Chowdhury, T. Xiang, and Y.-Z. Song, "Sketching without worrying: Noise-tolerant sketch-based image retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 999–1008.

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

[41] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deep generation of face images from sketches," *arXiv preprint arXiv:2006.01047*, 2020.

[42] Z. Huang, Y. Peng, T. Hibino, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "dualface: Two-stage drawing guidance for freehand portrait sketching," *Computational Visual Media*, vol. 8, no. 1, pp. 63–77, 2022.

[43] Z. Huang, H. Xie, T. Fukusato, and K. Miyata, "Anifacedrawing: Anime portrait exploration during your sketching," in *ACM SIGGRAPH 2023 conference proceedings*, 2023, pp. 1–11.

[44] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7915–7925.

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

[46] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuhori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer, "Streamdiffusion: A pipeline-level solution for real-time interactive generation," *arXiv preprint arXiv:2312.12491*, 2023.

[47] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[48] Pixiv Inc., "Pixiv," retrieved January 29, 2025. [Online]. Available: https://www.pixiv.net/

[49] PixAI Inc., "Pixai," retrieved January 29, 2025. [Online]. Available: https://pixai.art/

[50] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[51] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

Appendix

Appendix

In this appendix, we present the detailed statistics for the System Usability Scale (SUS) questionnaire items, interface preference questions, and component satisfaction ratings. Figures 7.2–7.11 display histograms of scores for SUS items 1–10; Figures 7.12–7.14 show the selection proportions for different interface preferences; and Figures 7.18–7.21 illustrate the distribution of user satisfaction scores for the individual components.

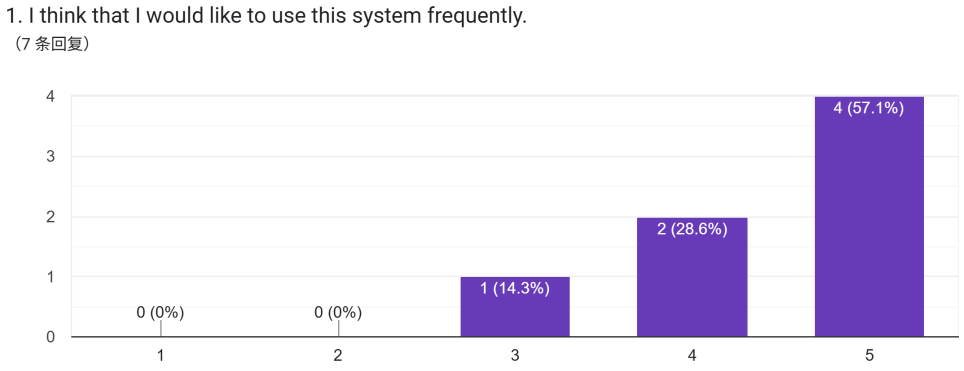


Figure 7.2: Distribution for SUS question 1.

2. I found the system unnecessarily complex.
（7 条回复）

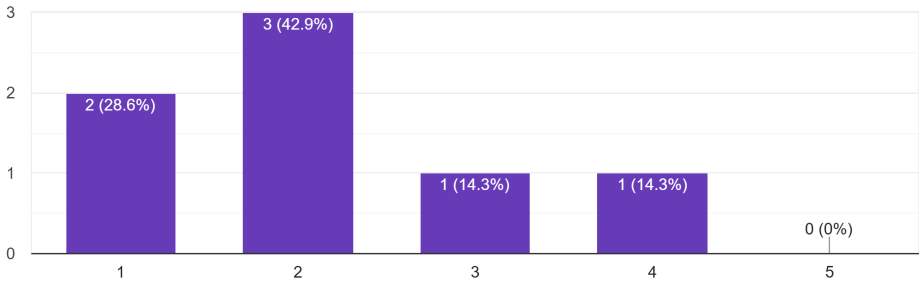


Figure 7.3: Distribution for SUS question 2.

3. I thought the system was easy to use.
（7 条回复）

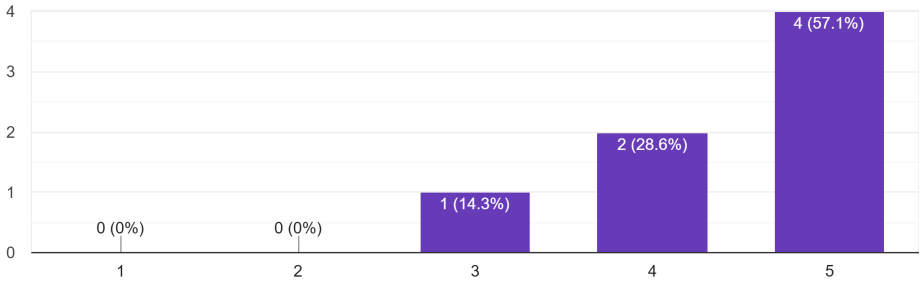


Figure 7.4: Distribution for SUS question 3.

4. I think that I would need the support of a technical person to be able to use this system.
（7 条回复）

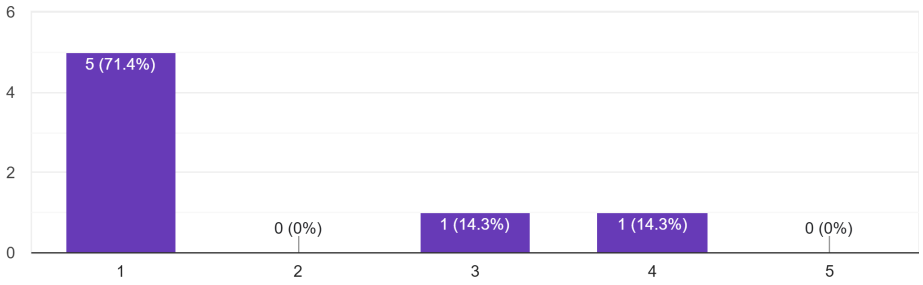


Figure 7.5: Distribution for SUS question 4.

5. I found the various functions in this system were well integrated.
（7 条回复）

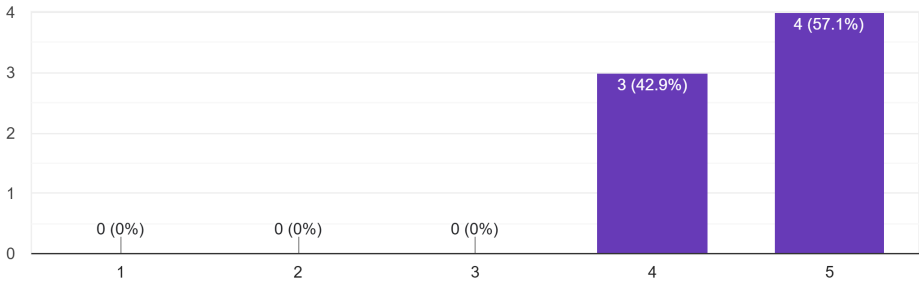


Figure 7.6: Distribution for SUS question 5.

6. I thought there was too much inconsistency in this system.
（7 条回复）

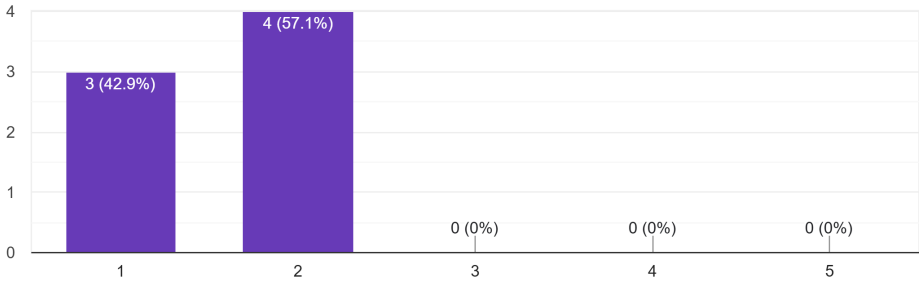


Figure 7.7: Distribution for SUS question 6.

7. I would imagine that most people would learn to use this system very quickly.
（7 条回复）

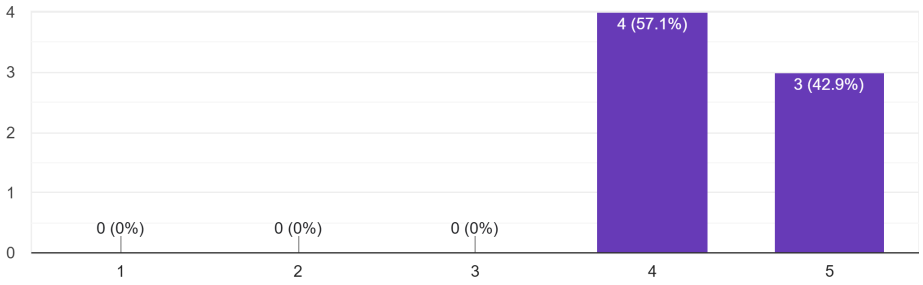


Figure 7.8: Distribution for SUS question 7.

8. I found the system very cumbersome to use.
（7 条回复）

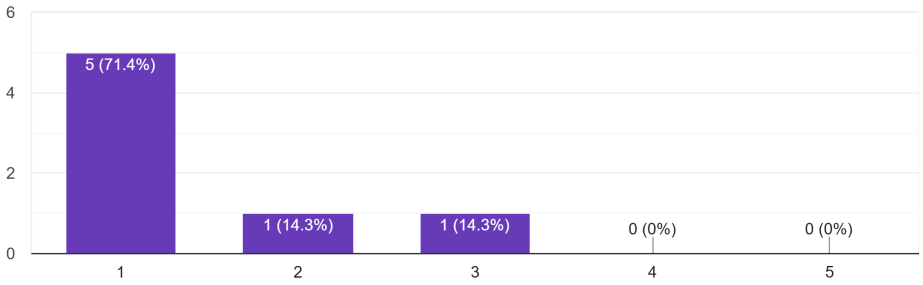


Figure 7.9: Distribution for SUS question 8.

9. I felt very confident using the system.
（7 条回复）

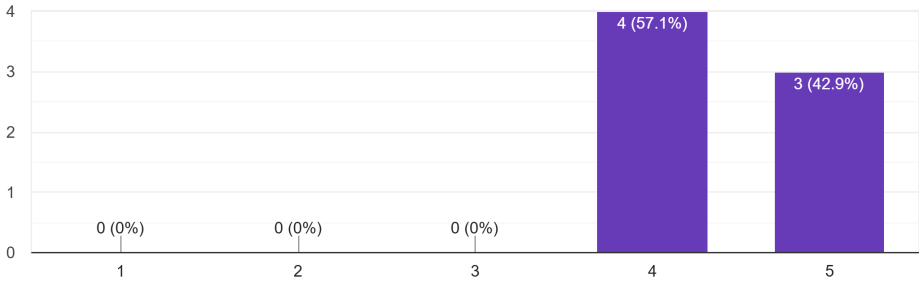


Figure 7.10: Distribution for SUS question 9.

10. I need to learn a lot of things before I could get going with this system.
（7 条回复）

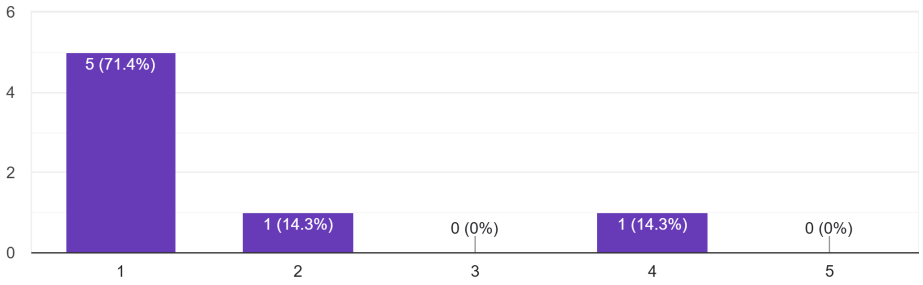


Figure 7.11: Distribution for SUS question 10.

11.Which of the three experiments do you think turned out to be the most consistent with what you had in mind when you started?

（7 条回复）

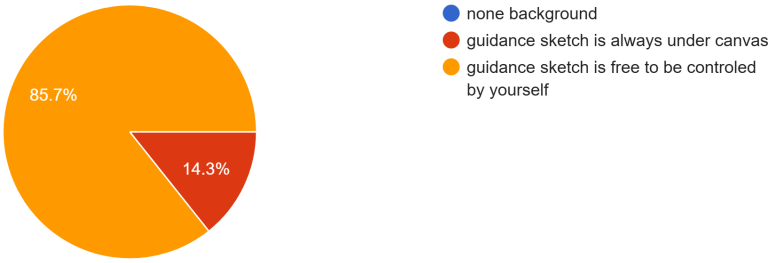


Figure 7.12: Distribution for preference question 11.

12.which is your desired outcome in the three groups

（7 条回复）

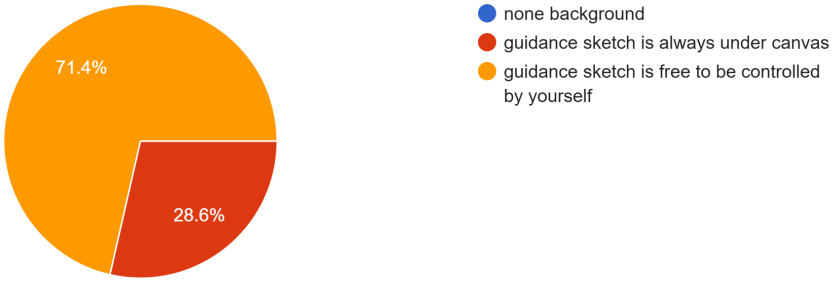


Figure 7.13: Distribution for preference question 12.

13.Which one do you think is better for you to draw

（7 条回复）

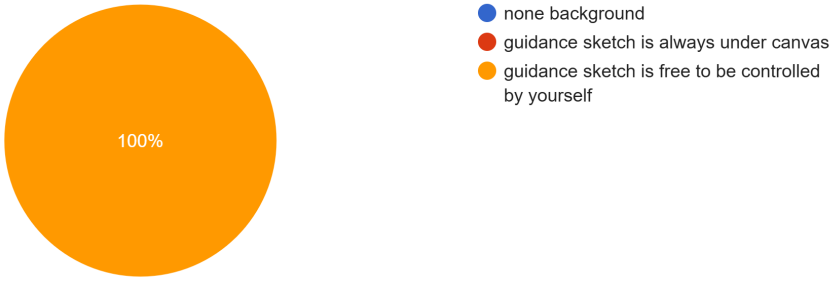


Figure 7.14: Distribution for preference question 13.

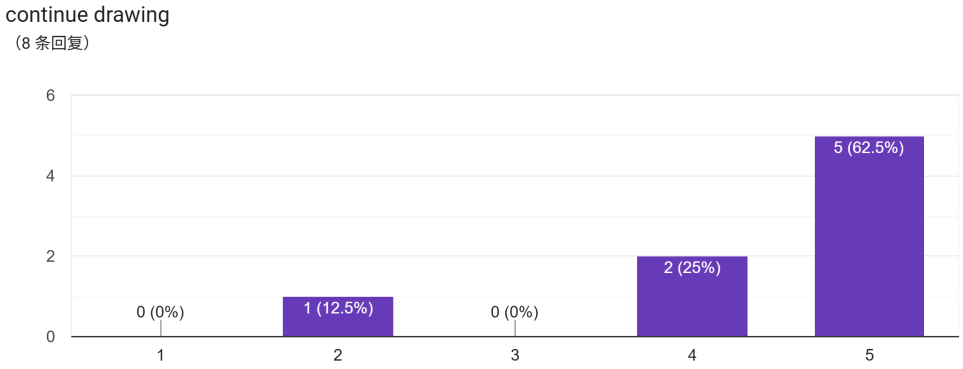


Figure 7.15: Satisfaction distribution for the “continue drawing” function.

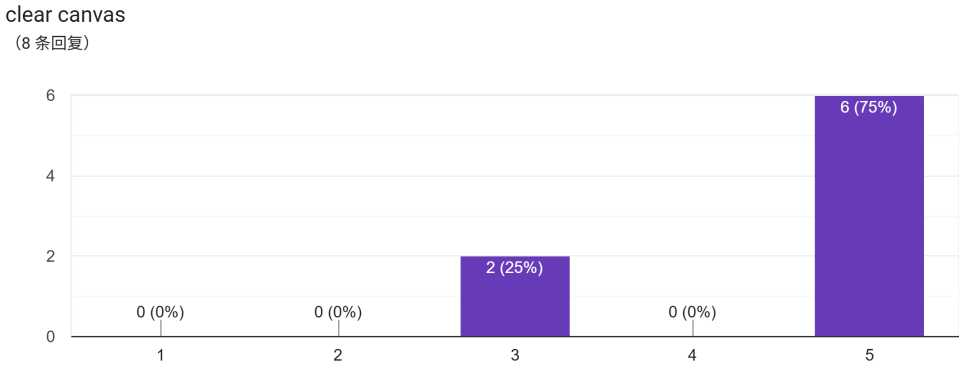


Figure 7.16: Satisfaction distribution for the“clear canvas” function.

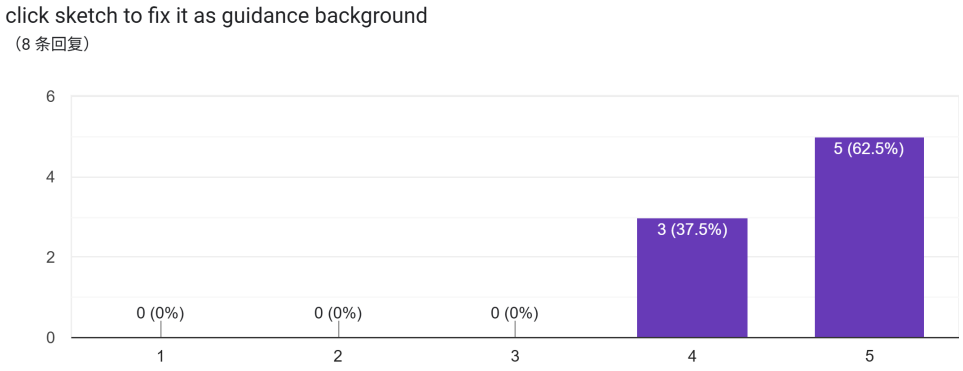


Figure 7.17: Satisfaction distribution for the “control the position of guidance sketches” function.

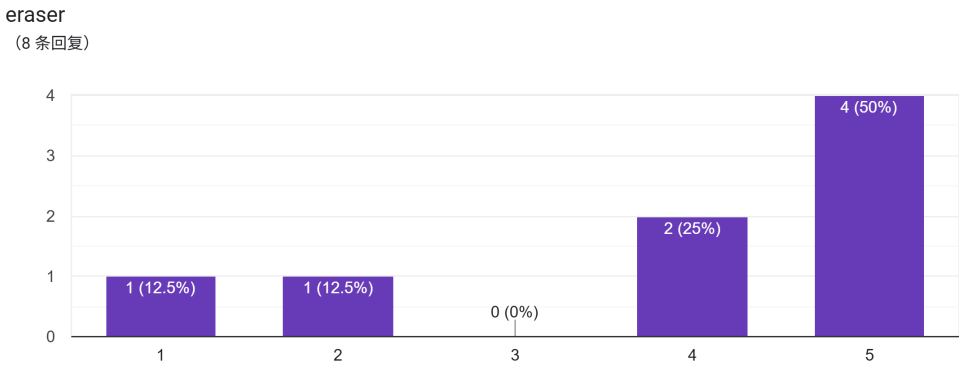


Figure 7.18: Satisfaction distribution for the “eraser” function.

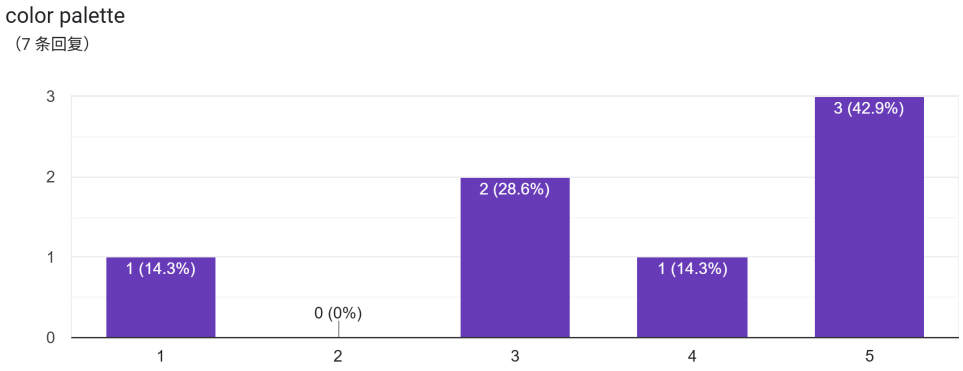


Figure 7.19: Satisfaction distribution for the“color palette” function.

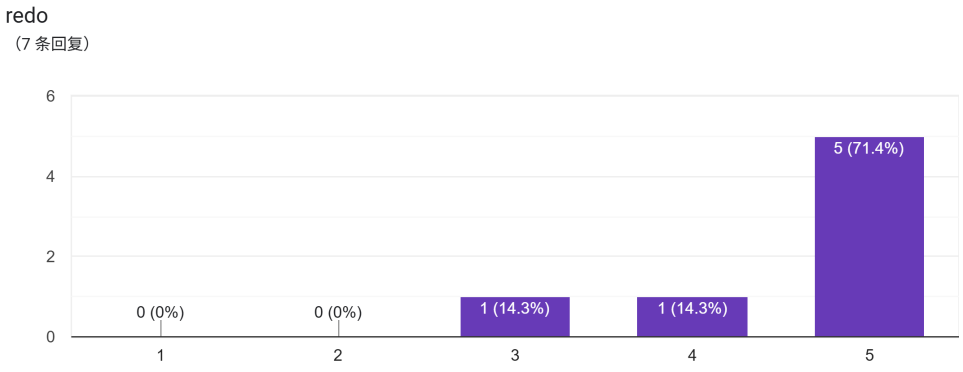


Figure 7.20: Satisfaction distribution for the “redo” function.

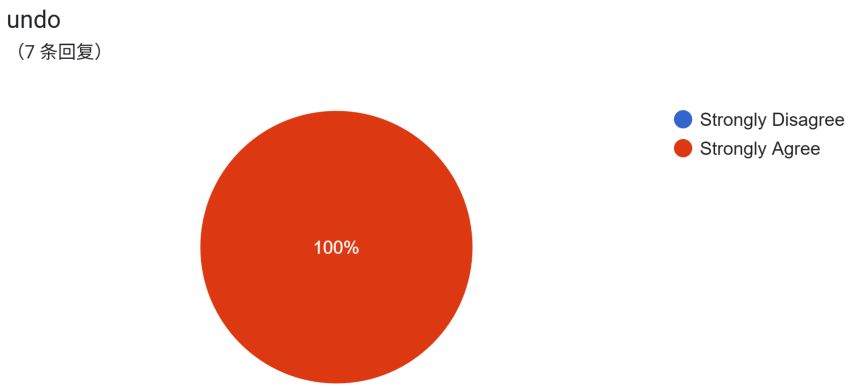


Figure 7.21: Satisfaction distribution for the “undo” function.