Doctoral Dissertation

CHARACTER DESIGN UNDER MULTIMODAL INPUTS BASED ON GENERATIVE MODELS

Sicheng Li

Supervisor Kazunori Miyata

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

March 2025

Abstract

In recent years, generative models, especially diffusion models, have made significant advancements, pushing the boundaries of traditional image synthesis and addressing various application-specific demands in fields such as virtual reality, film production, and fashion design. These technologies enable the automated generation of characters with precise specifications, dramatically improving workflow efficiency and reducing production costs.

One of the most transformative advancements in generative models is multimodal generation, which leverages diverse inputs like text prompts, image references, and spatial guidance (e.g., sketches or segmentation maps). This approach opens up broader creative possibilities, enabling flexible design workflows and improving model robustness and adaptability across tasks. As technology evolves, multimodal inputs are poised to revolutionize character design, enhancing both efficiency and creativity in various industries.

Building on multimodal generation, this dissertation investigates character creation and design through generative models incorporating diverse inputs like text, stroke data, and structural maps. The core challenge lies in balancing **manipulability**, **convenience**, and **precision**—three essential yet often competing elements in the design process:

- **Manipulability** refers to the level of control users have over the generative outputs, allowing flexibility in shaping design details.
- **Convenience** emphasizes the efficiency of the design workflow, streamlining the process to reduce repetitive tasks and focus on higher-level decisions.
- **Precision** indicates the model's ability to produce outputs that accurately align with user inputs, such as text descriptions or sketches.

Achieving an optimal balance among these factors is complex, as enhancing one often limits the others. This study addresses this **trilemma** by identifying the primary contradiction in each design scenario and resolving it according to user needs, guided by Marx's theory of contradictions. This approach enables tailored solutions across three design applications, each focusing on different primary tensions within the trilemma:

- **Drawing Multi-Age Facial Features for Anime Characters**: This task primarily balances **precision and manipulability**. Designing lively, age-specific facial features in anime characters demands both detailed control and real-time interactivity. I developed an inter-

active painting assistance system that leverages user-inputted strokes to create facial features with age-specific characteristics. This system ensures continuous interaction between the user's design intent and the generative model, effectively balancing user creativity and model-driven generation.

- **Character Pose Design**: This task requires balancing **precision and convenience**. Quickly generating character images that align with specific descriptions is crucial in fields such as advertising and poster design. Building upon conditional diffusion models like ControlNet, I developed an enhanced end-to-end text-to-image (T2I) generation framework that enables efficient, accurate pose creation. Users can generate custom character poses by combining text inputs with spatial conditions, such as skeletons, facial landmarks, and sketches, to produce diverse and precise poses.
- **Character Head Motion Design**: Here, the primary focus is balancing **convenience and manipulability** for video-based motion design. Traditional frame-by-frame animation is often labor-intensive, especially for designing head movements in character animation. I introduced a head motion prediction model, integrated into an image-to-video (I2V) generative framework, to streamline the workflow. This model uses multimodal inputs—trajectory strokes, audio, and reference images—to predict head movements, allowing for intuitive user control over the motion trajectory while reducing the manual effort required for animation.

In conclusion, this dissertation uniquely contributes to advancing multimodal generative models in character design, optimizing user-customizable workflows across varied scenarios. By carefully balancing manipulability, convenience, and precision, this research enhances the applicability of generative models in creative domains, fostering both efficiency and creativity.

Keywords: Character Design, Generative Models, Multimodal Generation, Computer Vision, Human-Computer Interaction.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The director of my favorite childhood animated film, Toy Story, John Lasseter, once said, "Art challenges technology, and technology inspires art." This quote perfectly captures the dynamic relationship between the evolution of generative models and human creativity. This paper delves into the core challenges that arise from the intersection of generative models and the character design process, employing a philosophical methodology to harness the full potential and adaptability of generative models in character creation.

1.1 Background and Significance

In recent years, generative models have made significant strides, transforming content creation and design workflows. Early models like Variational Autoencoders (VAEs) [4] laid the groundwork by learning latent representations, but often struggled with producing sharp, high-quality images. Generative Adversarial Networks (GANs) [5] further advanced the field by generating more realistic and detailed images through adversarial training, although challenges such as mode collapse and instability in training limited their effectiveness in complex tasks like character design. The most recent breakthrough has come with diffusion models [6], which generate highly detailed and consistent images by iteratively refining noisy data, making them particularly suited for tasks requiring precision and control, such as character creation.

A notable development in the field is the rise of cross-modal generation, exemplified by models like CLIP (Contrastive Language-Image Pre-Training) [7], which bridges the gap between textual and visual data. This advancement has fueled further progress in multimodal generation. In single-modal generative models, representation learning encodes information as numerical vectors or abstracts it into higher-level feature vectors. As shown in Figure 1.1, multimodal representation learning leverages the complementarity between different modalities, reducing redundancy, and enhancing the dimensionality of the generated object's representation. This enables the model to learn

more robust and enriched feature representations. This greatly expands the scope of creative possibilities, allowing users to combine multiple inputs in innovative ways, resulting in more detailed, accurate, and customizable outputs. In character design, these multimodal inputs make the creative process more intuitive and adaptable, enabling designers to achieve specific outcomes while maintaining creative flexibility and efficiency.

The evolution of these technologies is poised to radically transform human content creation in the near future, shifting from the previous experience-driven and data-driven methods to a generation-driven creative framework. As these technologies continue to reshape creative processes, it is crucial to explore how generative models and AI-driven systems can be effectively integrated into user workflows to address the opportunities and challenges posed by these advancements. In the subsequent sections of this chapter, I will delve deeper into this issue and outline our research objectives and methodology.

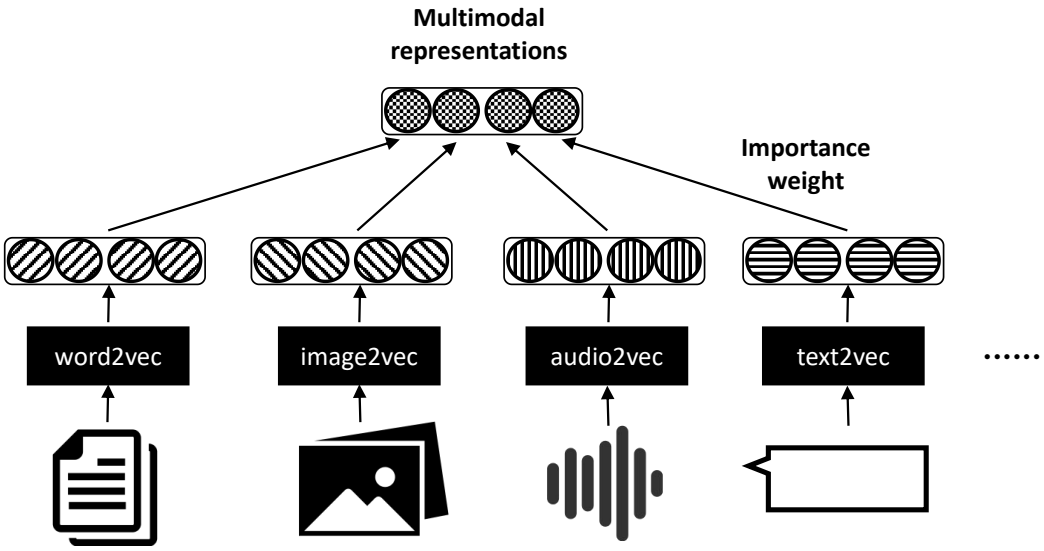


Figure 1.1: Multimodal representation learning through complementary modalities.

1.2 Motivation

Traditional character design workflows are often time-consuming and require extensive manual effort, with tasks like creating distinct facial features, designing expressive poses, and animating character movements demanding

significant labor and expertise. Designers typically go through multiple drafts, adjusting proportions and expressions, which can be inefficient. The integration of generative models, such as VAEs, GANs, and diffusion models, offers significant benefits by automating many of these repetitive tasks. In particular, multimodal generative models can reduce manual workload by generating high-quality characters from various input types—sketches, text descriptions, or reference images—allowing designers to focus on higher-level creative decisions while the models handle the more technical aspects of character creation. However, as generative models are increasingly explored in various creative processes, a prominent challenge has emerged—the growing discrepancy between the generated outcomes and the designer's creative intent. This discrepancy highlights the need for more effective frameworks that bridge the gap between the technical capabilities of generative models and the creative intent of designers. As generative models become more prominent, it is crucial to strike a balance between automation and user control, ensuring that these advanced tools do not overshadow the designer's vision but instead complement and enhance it.

The motivation behind this thesis is to explore reasonable frameworks for integrating generative models into various character design workflows in today's era, where generative models are gaining prominence. The goal is to ensure that technological advancements are fully utilized in the creative process, rather than limiting the creative output of designers or the powerful capabilities of generative models. This requires a careful examination of the interaction between human creativity and machine-generated outputs, focusing on how these technologies can be structured to support, rather than hinder the creative flow.

1.3 Challenges

The foundation of this study is built upon a trilemma that characterizes the relationship between design and generative models, framed by three key elements: manipulability, convenience, and precision. This trilemma serves as a theoretical framework for understanding how to balance human creativity with automated generative processes in character design.

- **Manipulability** refers to the degree of control that designers maintain during the creative process. Generative models should empower designers by allowing flexible manipulation of outputs, ensuring that the creative vision is not overshadowed by automation.
- **Convenience** emphasizes the efficiency generative models bring to the design workflow. Automation should simplify the process, reducing

repetitive tasks and allowing designers to focus on higher-level decisions without becoming bogged down in technical details.

- **Precision** refers to the consistency between the outputs generated by the model and the user's creative intent. High precision ensures that the model consistently produces outputs that accurately capture the user's intent, such as faithfully reflecting input like text descriptions, sketches, or structural maps.

These three elements represent different but equally important aspects of the creative process. However, inherent conflicts and contradictions exist among them. Enhancing one element may inadvertently compromise the others, creating a challenging balance for integrating generative models into character design workflows.

The core challenge lies in how the introduction of automation in generative models can disrupt traditional creative workflows. For instance, increasing convenience through automation might reduce the level of control (manipulability) a designer has over the final output. Similarly, striving for high precision in the generated outputs could complicate the process, diminishing convenience or limiting the designer's ability to make spontaneous adjustments.

Recognizing these conflicts is crucial for developing generative models that effectively support the creative process without undermining any essential aspect. The trilemma underscores the need for a balanced approach that carefully considers how improvements in one area might impact the others.

1.4 Methodological Framework

To address the challenges posed by the trilemma, the study adopts a methodological approach focused on identifying and resolving the primary contradictions between these elements. According to dialectical materialism, particularly Karl Marx's theory of contradictions, every process or system contains internal contradictions, with one primary contradiction playing a decisive role.

By focusing on the primary contradiction in each specific scenario, I can develop strategies that effectively balance the elements of manipulability, convenience, and precision. To further elucidate these relationships, I employ a mathematical perspective based on Bayesian probability theory.

As illustrated in Equation 1.1, the Bayesian marginal likelihood function provides a framework for understanding the trade-offs among the three elements:

$$P(x \mid y) = \int P(x \mid y, \theta) P(\theta \mid y) d\theta \tag{1.1}$$

Here, $P(x \mid y, \theta)$ represents the conditional probability of generating x given user input y and model parameters θ, capturing the model's capacity for detailed control through adjustments to both y and θ.

$P(\theta \mid y)$ denotes the posterior distribution of θ conditioned on y, reflecting the model's level of automation. A concentrated distribution minimizes user input, enhancing convenience, while a broader distribution supports more user control, increasing manipulability.

$P(x \mid y)$ represents the probability of achieving accurate outputs under input y, encapsulating the overall precision of generated results, which is influenced by both $P(x \mid y, \theta)$ and $P(\theta \mid y)$.

The trade-offs among these elements are as follows:

1. **Enhancing Convenience and Precision**: By strengthening the posterior $P(\theta \mid y)$ directly through user input y, the model can infer θ automatically, thereby enhancing $P(x \mid y)$'s precision. However, this automation reduces designer control over θ, limiting manipulability in favor of higher convenience and precision.

2. **Enhancing Manipulability and Precision**: By allowing multiple interventions through y (indirectly) or adjusting θ directly, $P(x \mid y, \theta)$ more closely aligns with the designer's intent, increasing $P(x \mid y)$'s precision. This, however, adds complexity and increases user workload, thereby reducing convenience.

3. **Enhancing Manipulability and Convenience**: A loose distribution $P(\theta \mid y)$ combined with detailed user input can support manipulability but might sacrifice convenience. Alternatively, a more constrained $P(\theta \mid y)$ could simplify automation but lack robustness to nuanced inputs y. Balancing both manipulability and convenience often requires more streamlined inputs y and broader distributions for $P(\theta \mid y)$, which may compromise the precision of $P(x \mid y, \theta)$, leading to outputs that may not fully meet the designer's expectations.

By applying this mathematical perspective, I can see that improving two elements often impacts the third. The key is to identify which contradiction is primary in a given scenario and develop methods to address it without excessively compromising the other elements.

To operationalize this approach, I explore three distinct character design scenarios, each representing a different balance within the trilemma:

- Facial feature design for anime characters of different ages: Here, the primary contradiction lies between precision and manipulability. The

user's main need is for the generative model to continuously provide precise, real-time facial feature references aligned with their design intent. To address this, I employ an iterative generation process that allows for multiple feedback loops throughout the design, even if it comes at the cost of reduced convenience in a one-time generation.

- Pose design for character creation: In this scenario, the primary contradiction is between precision and convenience. Users aim to generate accurate character poses using simple prompts and abstract conditions such as skeletons and sketches. Text and spatial signals, serving as input conditions, reflect a more global focus on the target's pose accuracy rather than fine details. To ensure the consistency between user input and system output, as well as maintaining ease of use, I implemented an end-to-end multimodal generative model framework, optimized specifically for output accuracy.
- Head motion design for characters: In this scenario, the primary tension is between manipulability and convenience. This task involves editing tons of video frames, which can be highly time-consuming and labor-intensive for users. To simplify this, generative models allow users to create head movements using simple inputs (trajectory strokes & reference images). The focus is on generating motion that reasonably follows the trajectory, rather than on achieving precise motion details. The key challenge is balancing the convenience of automated motion generation with the level of control needed for users to design the motion. To address this, I developed a trajectory-based head motion prediction model, integrated into a state-of-the-art I2V framework.

The selection of these three tasks is primarily because each represents a distinct main contradiction within the trilemma and each is a common and representative design scenario in character design. These three scenarios—static facial features, dynamic poses, and animated head motions—are essential and representative in the character design process. Each addresses different aspects of precision, manipulability, and convenience in the trilemma. By focusing on these widely encountered design challenges, the study ensures that the generative models developed are highly relevant and effectively support the diverse needs of character designers.

By focusing on the primary contradictions and understanding the trade-offs through a mathematical lens, the study aims to establish a framework where generative models support and enhance the designer's creative process in the most reasonable and user-concerned way. By doing so, I can develop generative models that effectively align with the designers' needs, providing the right mix of manipulability, convenience, and precision tailored to each

specific design scenario.

Figures 1.2 illustrate the key perspectives of this paper. Guided by the methodology of resolving primary contradictions, I identified the core user demands in various character design scenarios and used this understanding to construct an appropriate generative model framework that supports the creative process. Through extensive experimentation, I validated the effectiveness of the workflows, benchmarking the performance of our designed models against SOTA frameworks. The results confirm that our pipeline effectively addresses core design needs and provides substantial support for users in their creative endeavors. From my perspective, the principles and methodologies proposed here are not only applicable to the tasks discussed in this paper but can also be extended to a wide range of generative model-driven creative scenarios. Therefore, I believe that this work offers valuable theoretical foundations and methodological insights for future research and exploration in the field.

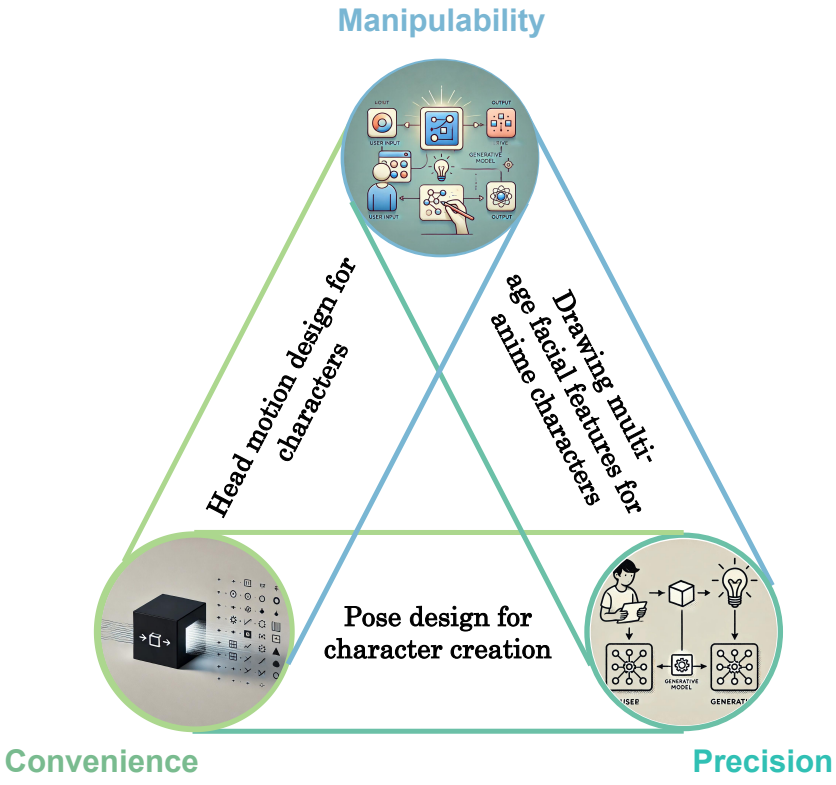


Figure 1.2: Primary contradictions in the trilemma across various design scenarios.

1.5 Organization of the Thesis

In Chapter 1, I discuss the influence of generative models on creative fields, introducing the "trilemma" of balancing manipulability, convenience, and precision in character design applications. Using the theory of primary contradictions, I identify key challenges in integrating generative models into design workflows. This study explores various design scenarios, demonstrating how tailored frameworks can effectively support designers' creative intentions while addressing inherent conflicts within the trilemma.

In Chapter 2, reviews fundamental theories and recent advancements in generative AI, with a focus on multimodal generation using diffusion models in human-centered design. By discussing key approaches, I establish the basis for our research and situate it within the current field of multi-modal generation.

In Chapter 3, I introduce a drawing support system aimed at assisting in the creation of anime characters with distinctive age-related facial features. This system, built on a Transformer generative model, encompasses its model architecture, drawing assistance strategies, and a specialized facial sketch dataset. A series of usability and effectiveness experiments confirm the system's ability to enhance character design.

In Chapter 4, I introduce our approach for more accurate Human Image Generation(HIG) conditioned on text and structure annotations, such as pose, landmarks, and sketches, which is particularly beneficial for applications like posters and advertisements. This chapter presents a Spatial Guidance Injector(SGI) and a Diffusion Consistency Loss(DCL) module designed to enhance pose accuracy within a text-to-image diffusion framework. Experimental results and ablation studies underscore the system's effectiveness in achieving accurate and consistent pose generation.

In Chapter 5, I focus on designing head movements for animated characters, introducing the Diff Transformer model to achieve trajectory-based head pose prediction, which is integrated into a diffusion model-based video generation framework. Our approach allows users to guide head pose dynamics using simple trajectories, resulting in realistic head movements. Comparative studies demonstrate the model's effectiveness in producing coherent and lifelike head animations.

In Chapter 6, I first validate the "trilemma" proposed in our thesis through a user study. Next, by summarizing quantitative and qualitative analyses from previous chapters, I highlight the effectiveness of our proposed strategies across diverse design tasks. Finally, I discuss the challenges encountered and limitations of the current models and propose future research

directions to advance the application of generative AI in creative character design.

Chapter 2

Related Works

Section 2.1 delves into the technical pathways and current research on multimodal generation, providing relevant examples and studies. In Section 2.2, I briefly outline the existing paradigms of generative models, analyzing their respective advantages, limitations, and differences. I then explore their impact on creative endeavors. In Section 2.3, I introduce applications and framework optimizations based on state-of-the-art diffusion models within user-centered creative fields. Finally, in Section 2.4, I described the challenges associated with generative models in design workflows from the perspective of artistic creation and, based on this foundation, proposed our theoretical framework.

2.1 Multimodal Generation with Transformers

Looking at a photo and describing it, or interpreting a complex scene and explaining its context, are relatively simple tasks for humans but can be significantly more challenging for computers. In recent years, however, numerous studies based on Transformer models have made remarkable progress in various multimodal tasks. Particularly, the success of large language models and their multimodal extensions [8–11] has further highlighted the potential of Transformers in multimodal generation tasks. Like other deep neural network architectures, Transformers also have substantial data requirements. With the introduction of increasingly large multimodal datasets, the combination of advanced models and multimodal big data has accelerated the development of multimodal generation techniques.

2.1.1 Multimodal Big Data

In the field of multimodal generation, especially for image and video synthesis, a diverse range of large-scale multimodal datasets has emerged, providing

critical resources for enhancing generative model capabilities across varied subjects.

The HumanArt dataset [12], for example, offers high-quality multimodal annotations for human faces and bodies, including pose keypoints, segmentation masks, and textual descriptions, making it highly useful for tasks in character design and stylized artistic generation. Datasets like FFHQ (Flickr-Faces-HQ) [13], although primarily focused on high-resolution facial images, provide a solid basis for face synthesis and editing with annotations on age, ethnicity, and expressions.

In video generation, AIST++ [14] combines 3D dance motion capture data with musical accompaniment, enabling synchronization between dance and music in generative tasks. Meanwhile, the YouTube-360 dataset [15] includes 360-degree video content with corresponding audio, which provides a valuable multimodal foundation for immersive and spatially aware generative tasks. Kinetics-700 [16] offers video clips paired with action labels, supporting action synthesis and human activity recognition across diverse scenarios.

The AVA (Atomic Visual Actions) dataset [17] includes multimodal data in the form of video clips annotated with action categories and temporal markers, targeting fine-grained generation and recognition of human actions in complex scenes. Finally, the Laion-5B dataset [18], with billions of image-text pairs, offers massive-scale multimodal resources that facilitate text-guided image generation across diverse visual themes, reinforcing cross-modal learning in generative models, MultiGen20M Dataset [19] provides more than 20M image-prompt-condition triplets. It includes 12 common control conditions(Canny, HED, Sketch, Depth, Normal, Skeleton, Bbox, Seg, Outpainting, Inpainting, Deblurring, Colorization).

These multimodal datasets provide not only the foundational data needed for generative tasks but also specialized multimodal annotations—like keypoints, segmentation, and synchronized audio-visual pairs—ensuring robust training for models across various creative domains.

2.1.2 Transformers

The Transformer architecture is a flexible and powerful framework that shares similarities with a generalized graph neural network. Its self-attention mechanism enables the processing of inputs by treating them as fully connected graphs, emphasizing global and non-local interactions. This characteristic allows the Transformer to handle diverse modalities in a modality-agnostic way, effectively representing the embedding of each token as a node in the graph.

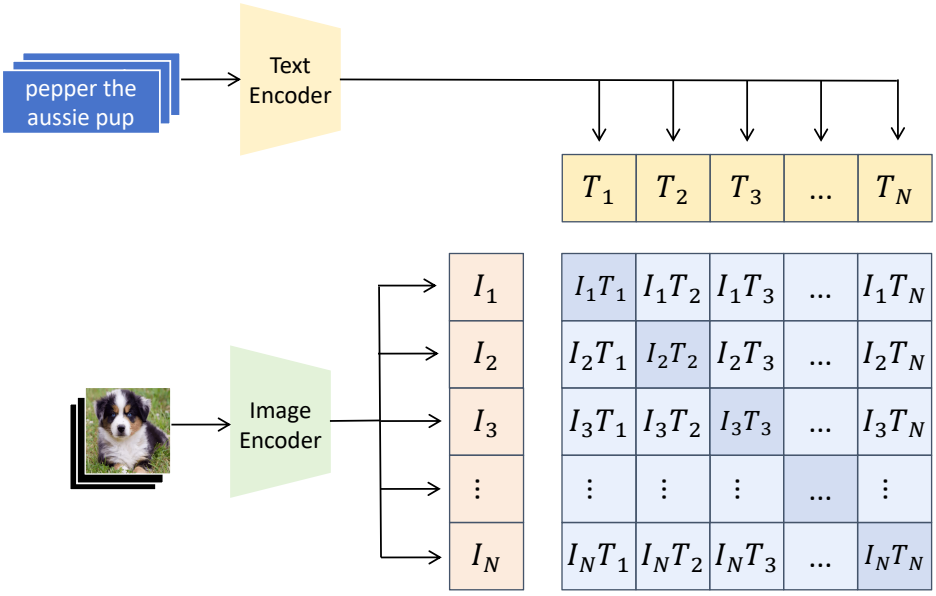


Figure 2.1: CLIP's shared embedding space for cross-modal text and image guidance.

Building on the success of Transformers, VideoBERT [20] pioneered the adaptation of Transformer architectures for multimodal applications, showcasing their extensive potential in multimodal domains. Following this breakthrough, several pre-training models based on Transformers—such as ViLBERT [21], LXMERT [22], VisualBERT [23], Pixel-BERT [24], Act-BERT [25], and ImageBERT [26] quickly became focal points in machine learning research. A significant milestone in this field was achieved with CLIP [27], which redefined multimodal pre-training by framing classification as a retrieval task, effectively enabling zero-shot recognition. CLIP leveraged large-scale multimodal pre-training to perform zero-shot learning successfully, demonstrating the robust application of Transformers in real-world tasks. Beyond CLIP, models like ALIGN [28] and Florence [29] have furthered the integration of Transformers in multimodal generation. ALIGN (A Large-scale Image and Noisy-text embedding model) extends CLIP's approach by training on vast, noisy datasets, improving model robustness in zero-shot image recognition and retrieval tasks. Its large-scale image-text pairing approach allows it to generalize effectively to diverse visual categories and concepts without specific fine-tuning. Florence expands on this by incorporating dense vision-specific modules that better align visual and textual information across more detailed semantic hierarchies, resulting in improved performance on complex multimodal tasks, including image

generation and scene understanding.

These Transformer-based frameworks excel not only in multimodal representation learning but also as key components in generative model applications. In models like Stable Diffusion, for example, CLIP embeddings play a critical role in conditioning the diffusion process to align generated images with the semantics of the text input. As illustrated in Figure 2.1 By mapping text and image modalities into a shared embedding space, CLIP enables seamless guidance of generative models through textual inputs, producing coherent and contextually relevant outputs. This integration of Transformer frameworks marks a major advancement in multimodal generative modeling, where self-attention mechanisms effectively learn from diverse data inputs. The capacity of these models to handle large multimodal datasets lays a robust foundation for sophisticated AI applications that demand nuanced cross-modal understanding and generation.

Leveraging the superior performance of Transformer models, they have been widely applied to various sketch-related tasks, such as sketch gestalt [30], sketch-to-image translation [31], and sketch-based retrieval [32]. However, these models often lack the cross-modal integration of pixel-level features and stroke mid-point coordinates when processing sketches, which is crucial for ensuring spatial similarity between the input and output. This spatial alignment is particularly necessary for facial sketch generation, as the output must maintain a certain degree of similarity to the input sketch. To address the above issues, I proposed AgeFace in Chapter 3, which enhances the alignment between generated sketches and the user's design intent.

2.2 Generative AI Models

Generative AI has consistently been a central area of inquiry within the field of artificial intelligence, and in recent years, various generative model paradigms have rapidly evolved. Models such as VAEs [4], GANs [5], flow-based models [33], and DMs [6] have greatly improved the quality of media generation across text, images, and video. The frameworks of these generative models are illustrated in Figure 2.2. VAEs (Variational Autoencoders) were among the first generative models, using a probabilistic encoder-decoder framework to learn latent representations of data [34]. They optimize a lower bound on the log-likelihood of data via variational inference, enabling efficient sampling from the latent space. GANs (Generative Adversarial Networks) introduced a game-theoretic approach to training, where a generator and discriminator are trained in opposition. This adversarial process allows the generator to learn data distributions and produce highly

realistic images. Flow-based models use invertible transformations and exact likelihood maximization, enabling efficient sampling and exact density estimation of data distributions [35]. DMs (Diffusion Models) leverage iterative denoising processes based on Markov chains to generate samples from noise. Their ability to progressively refine noisy data into high-quality samples has positioned them as state-of-the-art in image generation [36, 37].

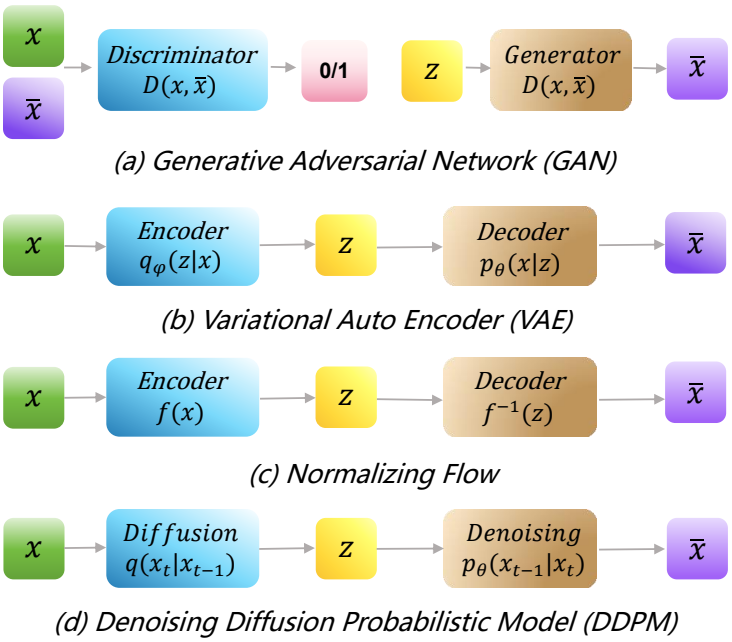


Figure 2.2: Generative model frameworks.

2.2.1 Diffusion Model

Most state-of-the-art diffusion models (DMs) are based on the Denoising Diffusion Probabilistic Model (DDPM), which constructs the image generation process through an iterative denoising sequence formulated within a probabilistic framework. In DDPM, a data point x_0 is gradually corrupted by adding Gaussian noise across a sequence of time steps $t = 1, \ldots, T$, resulting in a noisy latent representation x_T close to pure noise. The reverse process, defined as $p_\theta(x_{t-1} \mid x_t)$, iteratively denoises this representation back to x_0, reconstructing the original data. This reverse diffusion process is defined by the posterior probability:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}),$$

where $\mu_\theta(x_t, t)$ and σ_t represent learnable parameters, optimized to align the model's generative path with the observed data distribution. Through this iterative process, DDPM captures high-dimensional, complex distributions with remarkable fidelity, albeit at the cost of computational efficiency due to its multi-step generation process.

Building on this foundation, Latent Diffusion Models (LDMs) [37] operate within a perceptually compressed latent space, reducing dimensionality by projecting data into a lower-dimensional latent space, thus significantly enhancing sampling efficiency. In the LDM framework, the diffusion and denoising steps occur in this latent space, and the model decodes the latent representation back to the original data space for output, effectively reconstructing the high-quality image while retaining computational efficiency. The generation process in LDM can be expressed as follows:

$$p_\theta(z_{t-1} \mid z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \sigma_t^2 \mathbf{I}),$$

Where z represents the latent space variables, enabling efficient generation while maintaining high fidelity. Expanding from this base, DMs have seen development in several conditional diffusion models that enhance control and precision during generation.

2.3 Diffusion Models in Human-Centered Design

With the rapid advancement of diffusion models, numerous conditionally guided diffusion models have emerged in recent years. Among these, text-guided generative models are particularly prevalent, enabling high-quality, customizable outputs based on textual prompts. Leveraging the powerful generative capabilities of diffusion models, task-specific models have been developed for applications such as sketch-based, layout-based, and other visual cue-based generation, broadening diffusion models' applicability in targeted creative tasks. Additionally, multimodal-guided image/video editing techniques using text-to-image (T2I) diffusion models have become widely adopted tools for enhancing editing performance by allowing content modification guided by multimodal inputs. Together, these multimodal generation and editing technologies offer users a versatile design space for various creative scenarios.

2.3.1 Classifier Guidance

The Denoising Diffusion Probabilistic Model (DDPM) and its variants have demonstrated strong capabilities in generating realistic images. Building on these foundations, [38] introduced explicit classifier guidance, which enhances model control by injecting class label information during the generation process to enable conditional image synthesis. The key to diffusion models is the denoising step, where the probability $p(x_{t-1} \mid x_t)$ is modeled by a neural network to reconstruct data at each time step t. Classifier guidance modifies this by transforming the original $p(x_{t-1} \mid x_t)$ in an unconditional model into a conditional version, $p(x_{t-1} \mid x_t, y)$, where y represents the given label, as shown in the equation below:

$$p(x_{t-1} \mid x_t, y) = \frac{p(y \mid x_{t-1})p(x_{t-1} \mid x_t)}{p(y \mid x_t)}$$

This approach applies Bayes' rule, effectively allowing an unconditional model to produce conditional outputs without retraining the entire model. While classifier guidance is computationally efficient, it introduces some limitations: (1) An additional classifier must be trained, increasing the model's overall complexity. (2) The classifier performance influences the quality of generation, yet optimizing it can be challenging. (3) Classifier guidance may reduce the diversity of generated results by constraining the model to specific class-driven outputs.

To address these issues, [39] proposed classifier-free guidance. Unlike explicit classifier guidance, classifier-free guidance defines $p(x_{t-1} \mid x_t, y)$ as a Gaussian distribution:

$$p(x_{t-1} \mid x_t, y) = \mathcal{N}(x_{t-1}; \mu(x_t, y), \sigma_t^2 \mathbf{I}),$$

Allowing conditional generation by training both conditional and unconditional models. During training, the conditional input y is randomly dropped (replaced with null values), which simplifies training while increasing flexibility. This technique has been widely adopted in large-scale image generation models such as GLIDE, Stable Diffusion, DALLE-2, and Imagen due to its remarkable performance and adaptability.

2.3.2 Text-Guided Diffusion Models

Diffusion models have shown substantial potential in text-to-image generation, leading to the development of large-scale models like GLIDE [40], Stable Diffusion [37], DALLE-2 [41], and Imagen [42]. These models achieve

high image quality by integrating advanced techniques and progressively increasing parameter sizes.

GLIDE was one of the pioneering models in this space, using a cascaded architecture with classifier-free guidance. It generates an initial 64x64 image based on textual input and then upscales it to 256x256 with a text-conditioned super-resolution model, thereby improving output quality and alignment with prompts.

Stable Diffusion addresses the high computational cost typical of diffusion models by conducting diffusion in a latent space rather than directly on images. This method uses an encoder-decoder pair for perceptual compression, encoding text conditions into vectors with a text encoder. Cross-attention layers map these vectors into the latent space, streamlining the process and reducing computational demands.

DALLE-2 follows a two-stage approach. It first encodes text descriptions using CLIP (Contrastive Language-Image Pre-Training) [7] and generates corresponding image encodings with an autoregressive or diffusion model. Then, a decoder produces the final image, gradually upscaling it from an initial 64x64 to a high-resolution 1024x1024 image through super-resolution. Unlike typical diffusion models, DALLE-2 focuses on image features rather than noise prediction in its reverse process, yielding highly detailed results.

In contrast, Imagen adopts a single-step approach, favoring simplicity and high performance. Rather than modifying U-Net structures, Imagen improves text-to-image quality by using a powerful 11-billion-parameter text encoder, highlighting the importance of robust text representation in achieving superior image quality.

2.3.3 Subject-Guided Diffusion Models

While existing large-scale text-to-image models generate highly realistic and diverse images, they often struggle to create nuanced variations of subjects within reference images. To address this, IP-Adapter [43] facilitates image editing by fine-tuning keys and values in the cross-attention layers of pre-trained text-image generation models. This approach enables fine-grained control based on specific text prompts without altering the essential properties of the subject.

DreamBooth [44] enables synthesizing new images of a given subject in varied contexts from only 3-5 subject images and a text prompt. Its core approach links the subject and its identifier within a pre-trained text-to-image model, mapping this relationship to the output domain. This method maintains high fidelity, even for complex subjects, such as animals or specific objects.

LoRA (Low-Rank Adaptation) [45] takes a different approach, achieving flexibility across multiple tasks by adding low-rank matrices to specific layers of a model, which adapt without requiring full re-training. This adaptation mechanism offers efficient customization for varied downstream tasks.

However, both IP-Adapter and DreamBooth face limitations in diversity and controllability. DreamArtist [46] addresses these constraints by introducing a dual learning strategy in which the model learns expressive latent vectors from both forward and reverse processes using a text encoder and denoising network. This strategy improves both feature retention from reference images and output controllability, enhancing detail and diversity in generated images.

2.3.4 Sketch-Guided Diffusion Models

Conditional image generation based on diffusion models has proven effective in producing images with notable diversity and realism. However, most existing methods limit control over the final output to adjustments of labels or text prompts, thus constraining the degree of customization. In response, several sketch-based conditional image generation methods have emerged [47–51].

PITI [47] builds upon Glide, leveraging image layouts or sketches as input conditions. This model maps the input condition to the latent space of a pretrained model, which is then decoded to generate the final image. Sketch-Guided Diffusion [48], on the other hand, directly uses sketches to guide a pretrained text-to-image generation model, bypassing the need for retraining. The key innovation here is the introduction of a trainable latent vector predictor, based on a multilayer perceptron, which maps the latent features of a noisy image to a spatial map. Trained over thousands of images, this predictor operates on each latent pixel, offering flexibility and adaptability.

Generating high-quality facial images from sketches introduces a unique challenge: the need to construct high-dimensional facial features while preserving the visual detail within the sketch. Many current models treat sketches as auxiliary information, guiding the generation process but often sacrificing critical sketch details. DiffFaceSketch [51] addresses this by using sketches as the sole input, training in two stages for both sketch encoding and image generation. It also employs data augmentation techniques to synthesize varying degrees of facial abstraction from the input sketch, ensuring that sketch details are accurately preserved and effectively translated into the final output.

2.3.5 Multi-Condition Guided Diffusion Models

With the development of diffusion models, the limitations of using single or restricted conditioning methods have become increasingly apparent. To improve the practicality and controllability of image generation, various conditional guidance forms have been widely adopted. ControlNet [52], for instance, guides image generation with multiple conditions by locking the parameters of a pre-trained Stable Diffusion network and then duplicating the locked network to incorporate conditioning information. This process effectively fine-tunes the pre-trained network, enabling high-quality image generation based on detailed edge maps, abstract sketches, human poses, and more.

Composer [53] further expands the range of input conditions, including text descriptions, depth maps, sketches, color maps, style references, and masks, allowing for refined control over image generation. This model effectively integrates local and global information, supporting tasks like style transfer and other image transformation operations.

To more seamlessly combine the internal knowledge of pre-trained models with external control signals, T2I-Adapter [54] explores the use of lightweight adapter models. This approach introduces plug-and-play adapters that, while minimally impacting the original network structure, can flexibly combine different types of conditioning inputs. T2I-Adapter is therefore highly adaptable and efficient, offering versatile generation options through a range of composable conditions.

Although diffusion-based controllable frameworks have achieved remarkable progress recently, they still face limitations in control precision due to inherent conflicts between text and image control. The precision issue often requires multiple iterative generations in practical design workflows, increasing process complexity. To address these challenges, I proposed ECNet in Chapter 4, which significantly improves the accuracy of controllable SD-based models and enhances the application's convenience.

2.3.6 Diffusion Models for Editing

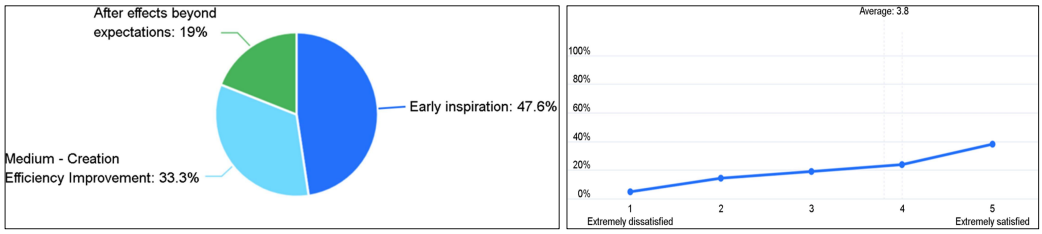
For image editing tasks, instruction-based editing built upon T2I models provides an intuitive approach for human image manipulation, where users input command-style text instead of detailed descriptions. Researchers [55–57] emphasize the necessity of collecting sufficient training triplets comprising instructions, source images, and corresponding edited images. InstructPix2Pix [55] optimized this approach by leveraging GPT-3 [58] to enable zero-shot editing, transforming images based on source and target descriptions.

Additionally, some studies [59, 60] have incorporated traditional visual tasks into instruction-based editing frameworks. For instance, InstructDiffusion [59] introduced the IEIW (Image Editing in the Wild) dataset to unify diverse tasks such as segmentation and keypoint detection. SmartEdit [61], tackling more complex challenges, introduced a Bidirectional Interaction Module (BIM) designed to process image features extracted from the LLaVA visual encoder [62], which incorporates detailed information essential for refined visual transformations.

Integrating objects from reference images into a source image is another challenging task, requiring a coherent composition of distinct elements. Paint-by-Example (PbE) [63] achieves this by using a CLIP encoder to extract the global semantics of the reference image, which are subsequently aligned within the base model using cross-attention layers. ObjectStitch [64] enhances this process by introducing a content adapter that processes and fuses reference images with the source, ensuring coherence during each denoising step to retain background content. Additionally, Reference-based Image Composition (RIC) [65] utilizes sketch-based structural control within masked regions, enabling improved compositional alignment.

For video editing tasks, the challenge of temporal inconsistency arises because T2I models are trained on static images, and editing each frame individually often results in inconsistencies. To address temporal coherence, some studies utilize pre-trained video diffusion models (VDMs) [66, 67], while others train video editing models from scratch [68, 69]. Works like MagicVideoEV [67] and AnyV2V [70] use pre-trained VDMs to guide motion consistency. Dreamix [71], for example, feeds scaled-down noise videos along with text prompts into a cascaded VDM [66], yielding temporally consistent edited videos. Meanwhile, AnyV2V edits the first frame and generates the entire video through an image-to-video model [72].

Current diffusion-based video editing frameworks primarily focus on the global motion of the target while overlooking the natural coherence of accompanying local motions. This limitation often results in unnatural inconsistencies within the overall motion, diminishing the rationality of the design outcomes. To address this issue, I proposed a novel framework in Chapter 5 that focuses on generating natural and coherent character head motion videos based on trajectories. This framework effectively resolves the shortcomings of current methods, ensuring improved rationality and consistency in design outcomes.



(a) (b)

Figure 2.3: User insights on the AI-assisted character creation process. Subfigure (a) illustrates the proportion of users employing AI assistance at different stages of the creative process, while Subfigure (b) presents user satisfaction levels with AI-assisted design, the figure is from [1].

2.4 Creation With Generative AI

Artistic creation is inseparable from the innovation of human thoughts, emotions, and inspiration. When AI painting becomes a method with broad application prospects and value, people will adjust their creative practices and languages [73].

With rapid advancements in AI technology, generative models have become deeply integrated into modern art and design, providing users with fresh inspiration, reducing repetitive tasks, and fostering a seamless fusion between art and technology. In today's creative workflows, AI tools primarily serve as supportive assets, assisting artists in generating images while leaving the core inspiration and decision-making to the human creator. However, most existing research focuses on the technical effectiveness of AI-generated outputs, with less attention paid to how these tools impact the broader design process. Therefore, a more in-depth understanding of character creation with AI tools—and how these tools align with various design goals-is crucial to advancing this field. Identifying these distinctions allows us to enhance efficiency and quality across industries and tap into new commercial possibilities.

Prior research has shown that AI systems, such as text-based models like ChatGPT, can be leveraged for imaginative art creation, with their extensive artistic knowledge contributing to diverse forms of creative generation [74]. Additionally, studies have examined the limitations and challenges of tools like ChatGPT and AI-based art generation in design, exploring generative AI's potential impact on creative processes and the development of design systems [75]. Simultaneously, other scholars have been investigating the use of AI image generation techniques in personalized cultural product design, presenting new avenues for creativity in this domain [76]. In another study,

researchers conducted in-depth surveys with 14 designers from various fields, analyzing AI-based character creation methods and gathering insights on diverse user needs for generative AI, as illustrated in Figure 2.3. These studies highlight the varying demands users have for generative AI, forming the basis of our proposed "trilemma" theory.

Chapter 3

Drawing Multi-Age Facial Features for Anime Characters

Drawing anime characters with facial features of different ages is a challenging task. The characters' facial features vary significantly with age, making it especially difficult for beginners to depict age-specific anime characters accurately. This task highlights the critical contradiction between **Precision** and **Manipulability**: designers require precise, age-specific facial features while retaining control over artistic adjustments. This scenario exemplifies character design in the context of customization.

To address these challenges, I propose AgeFace, a drawing support system that integrates interactive drawing guidance with generative models to help users balance the demands of precision and manipulability throughout the creative process. AgeFace can provide a combination of local and global user guidance in the drawing process to enhance both detailed facial features and overall aging features. Local guidance assists users in drawing detailed facial features, while global guidance provides hints for the overall layout of the face and additional features, such as wrinkles. During the local guidance stage, I apply an image retrieval approach to provide detailed instructions on facial features. In the global guidance stage, I propose the Transformer-based sequential generation model to create entire anime faces from drawn stroke sequences. The proposed framework of AgeFace combines a data-driven retrieval method and the generation model to provide users with inspiration during the drawing process. To verify the effectiveness of our guidance, I conducted user studies and comparison experiments with existing sketch generation models. The results demonstrated that AgeFace can significantly help users create multi-age anime faces and validate the effectiveness of our proposed generative model.

3.1 Introduction

Drawing facial age features is an important part of creating anime characters. However, the current process of creating facial features at different ages mainly relies on users' personal drawing experiences. As a result, it is challenging for novices to accurately and meticulously draw each facial feature of different ages. Additionally, it has been observed that users without sufficient experience in drawing anime characters, even those with basic drawing skills, find it challenging to create satisfactory age features. Prior work in data-driven support for drawing dynamically provides drawing guidance for users, such as ShadowDraw [77]. This approach furnishes the user with comprehensive information regarding the subject matter throughout the drawing process. While this may enhance the user's drawing proficiency, it is likely to alter the user's original design intentions. In addition, due to the limited number of references in the database, relying solely on database retrieval for drawing guidance can restrict users' creative diversity. Therefore, I aim to develop an advanced drawing support system that combines a generative model with data-driven methods to provide both local and global guidance, helping users accurately create facial features for different ages while preserving their original design intentions. Additionally, the lack of a database for the facial features of anime characters across different age groups makes it challenging to implement data-driven methods to support the drawing process.

To address the above concerns, I propose a user interface (UI) designed to support users in drawing facial features for multiple ages. To this end, I have developed a stroke-based sketch dataset comprising age-specific freehand facial sketches. This dataset aids users in producing high-quality facial representations for three distinct age categories and genders: male/female child, male/female middle-aged(including young and adult), and male/female elderly. I observed significant differences in facial features when drawing children(around 5-12 years old), middle-aged adults(around 16-45 years old), and elderly characters(above 60 years old) in anime. Within the middle-aged category, as shown in Figure 3.1, while youth and adult characters have differences in detail, they are not as pronounced as the differences between children and the elderly. Therefore, to simplify classification, I combined youth and middle-aged into a single group.

Furthermore, to better uphold the user's original design intentions, I developed an innovative drawing guidance strategy. This strategy integrates feature-focused guidance, which I call "local guidance", with traditional guidance based on all facial features, which I call "global guidance".

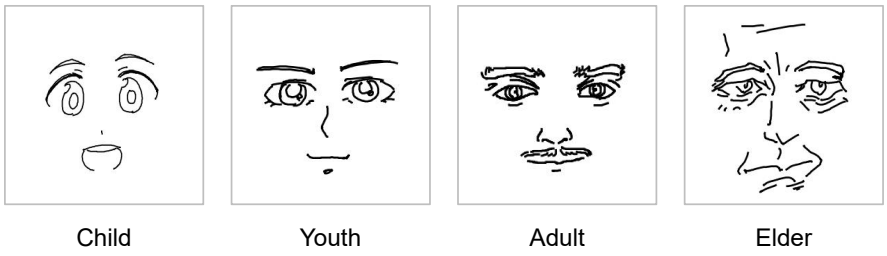


Figure 3.1: Differences between anime faces of four different age groups.

To address the constraints imposed on drawing results by database retrieval, I utilized a collected dataset to train a generative model, thereby enhancing the diversity of the drawn sketches. Considering that sketch images are sparse matrices, treating sketches as stroke sequences rather than pixel-based representations increases the stability of the generated results. Therefore, I propose a sketch sequence generation model using the Transformer architecture. This model encodes stroke sequences into tokens to learn deep representations of the complex structures in face sketches. Simultaneously, to minimize discrepancies between the generated results and the user's input sketches, I incorporated a ResNet module within the Transformer structure to extract image features. The structural information extracted by the convolutional layers allows the generated sketches to retain identity information from the input sketches and improves the robustness of their spatial structure.

Using our multi-age anime face dataset, I combine retrieval and generation methods to guide the drawing process. Specifically, during the local guidance stage, I retrieve matching facial features from the dataset for drawing assistance. During the global guidance stage, I offer both retrieved and generated whole-face sketches, allowing users to select their preferred candidate as the global guide. This approach enables users to complete the drawing process with combined local and global guidance, the framework of AgeFace is shown in Figure 3.2.

The main contributions of this research are as follows:

- I developed a UI to establish a database that collects freehand sketches, including stroke information. This interface features an interactive system that enables users to quickly categorize strokes based on facial features, such as eyes and noses. I hired expert designers to utilize this interface, constructing a stroke-based sketch database comprising 120 freehand sketches of male and female subjects across three distinct age groups (with 20 sketches per gender/age group).

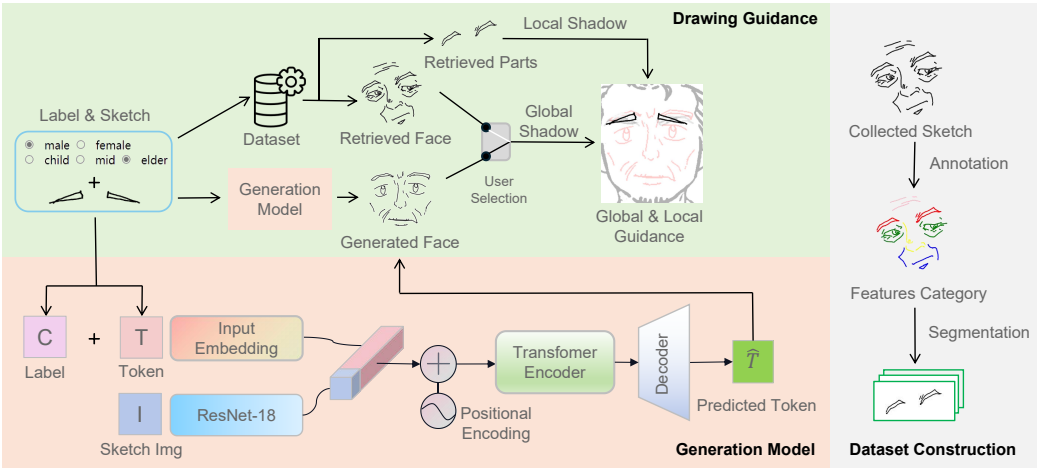


Figure 3.2: Framework of the proposed AgeFace, which includes three parts: data construction, drawing guidance interface, and Transformer-based generation model.

- Building on previous drawing guidance methodologies, I propose a novel strategy that combines global guidance, which encompasses all facial features, with local guidance focused on individual characteristics, thereby better respecting the user's design intentions during the guidance process.
- To enhance the diversity of drawing guidance, I designed a generative model based on the Transformer architecture, trained using our collected database. I integrated this model into the drawing guidance system to provide users with richer guidance and further inspire their creative process.

3.2 Related Works

3.2.1 Data-Driven Drawing Support System

Many data-driven interaction systems have already been introduced for sketching and drawing [78, 79]. For example, ShadowDraw [77] and Sketch Helper [80] search candidate results based on real-time drawing strokes to generate shadow guidance for users. DrawFromDrawings [81] assists users with 2D drawings by providing the retrieved sketch image from its dataset as a reference. DualFace [82] provides a human-AI co-creative drawing system that proposes two-stage drawing guidance to assist the user in completing

better freehand portrait sketches. However, the references provided by the systems mentioned above are based on the overall image. These drawing assistance systems provide complete visual guidance by displaying entire facial features based on retrieved or pre-generated references. While these systems enhance drawing accuracy, they often override users' creative intent by imposing fixed design patterns, ultimately compromising Precision in terms of user-driven artistic control.

Therefore, to preserve the user's original creative intent, I provide the user with a partial guide to the currently drawn facial features by identifying the facial feature attribute of the current stroke.

3.2.2 Sketch datasets of faces

To accommodate the growing demands of deep learning, numerous large-scale sketch datasets have been assembled. The Quick, Draw! dataset [83], for example, comprises over 50 million freehand sketches spanning 345 categories. Sketches within this dataset were amassed by instructing users to draw specified objects within a brief time frame. Consequently, the face subcategory, which includes 148,436 drawings, mainly exhibits oversimplified facial features with restricted expressions. FaceX [84] accumulated a vast dataset consisting of over 200,000 face sketches by amalgamating thousands of sketches of facial features drawn by adept designers. The CUFS dataset [85] collected 606 pixel-based portrait sketches, which artists created based on neutral expression photographs taken in a frontal pose.

However, no currently available dataset includes face sketches of varying ages, particularly in vector format. Furthermore, extracting line drawings from existing face databases, such as the Anime Face Dataset [86] and CelebA(a face attributes dataset) [87], as drawing references pose significant challenges. The resulting line drawings are ill-suited for use as drawing references due to an excessive number of lines. Some works [88, 89] have managed to extract sketches from images, but the results tend to be simple line drawings rather than detailed and refined sketches.

To address the lack of an aging face sketch dataset, I recruited a group of expert designers—comprising graduate students specializing in fine arts and seasoned illustrators—to create facial sketches. These sketches were then utilized to construct a stroke-based dataset. This dataset meets our needs for retrieval-based drawing support and training generative models, filling the gap in aging face sketch datasets.

3.2.3 sketch-based generative models

Using deep learning models for sketch generation has long been a popular research direction. Early research, such as DoodleGAN [90], treated sketch strokes as pixel information and used generative adversarial networks (GANs) for image generation. Thanks to the impressive generative capabilities of diffusion models [37, 91–93], some interesting projects [94] have also attempted to generate sketches from images. However, since sketches differ from typical RGB images or photographs in their sparse data and lack of color information, traditional image generation methods often fail to produce high-quality sketches. This inherent characteristic of sketches necessitates exploring alternative generation models, moving away from traditional pixel-based methods.

In recent years, researchers have proposed and developed sequential generation models specifically designed for sketch data. These models focus on the temporal properties of stroke sequences. By training on these stroke sequences, the models can learn the underlying structures and patterns in sketch data. Consequently, the generated sketches exhibit higher consistency and fidelity, closely resembling hand-drawn sketches.

Several variations of sequential generation models have been proposed [95–99], including those that leverage recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer architectures. However, for the specific task of face sketch generation, existing models still possess certain shortcomings. The LSTM-based model SketchRNN [100] exhibits limitations, such as insufficient sketch generation quality and inferior performance for sketches with longer strokes. Sketchformer [31] demonstrates that, for complex sketches with long stroke sequences, the Transformer architecture significantly improves reconstruction quality compared to LSTM structures. Nonetheless, existing Transformer models related to sketch data are primarily employed for tasks such as sketch gestalt, sketch-to-image translation, and sketch-based retrieval. These models do not incorporate pixel-level features of sketches to ensure spatial similarity between input and output, which is essential when employing generated results for face sketch drawing assistance, as the output face sketch must maintain a certain level of similarity to the input.

These generative models, including pixel-level and sequence generation frameworks, introduce automation into drawing tasks. However, designing facial features in anime requires Manipulability for precise artistic adjustments, which these highly automated models often fail to provide. Their black-box nature reduces designers' ability to control fine details, restricting the dynamic interaction needed for character manipulability. Our proposed

method addresses this contradiction through a Transformer-based stroke-sequence generation model that ensures precise age-specific facial features while maintaining interactive sketch adjustments. This dual-guidance approach resolves the Precision-Manipulability conflict by balancing accurate feature representation with real-time customization capabilities.

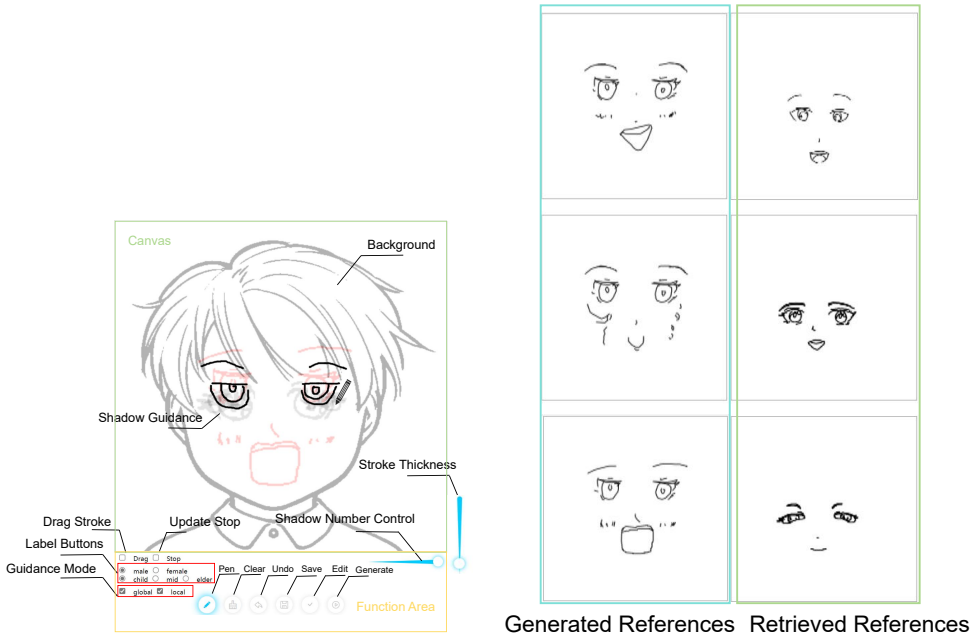
3.3 Proposed Method

In this section, I introduce AgeFace, our proposed drawing support system designed to assist users in creating anime faces of multiple ages. The system combines retrieval and generative methods to offer both local and global guidance, effectively enhancing users' drawing experiences and maintaining their original design intents. I explain the UI, dataset construction, generation model, and integration of shadow guidance to demonstrate the system's functionality and effectiveness.

3.3.1 User interface

The UI of our proposed system is divided into three primary sections: canvas, function area, and reference sub-windows. The canvas allows users to draw sketches, display background images, and view shadow guidance. The function area contains buttons for common actions, such as brushing and saving. The two areas are illustrated in Figure 3.3(a). The "Shadow Number Control" slider enables users to adjust the number of shadows displayed, with a range from 1 to 3. The "Stroke Thickness" slider allows for the adjustment of brush thickness. The "Drag" checkbox enables users to select and move strokes, while the "Stop" checkbox halts guidance updates. Label buttons facilitate switching between background images to accommodate different gender or age characteristics. Users can also choose between three navigation modes by checking the "Global" and "Local" boxes. The "Generate" button allows users to create high-quality facial feature sketches using the generative model.

To enrich the reference options, I provided six sub-windows on the right side of the drawing board, as shown in Figure 3.3(b). The three on the left display generated sketches, and the three on the right display retrieved sketches. This setup offers users a variety of reference choices. Users can select their preferred sketches as global guidance (highlighted with a red shadow) to assist with their drawing. Notably, users can enter stroke editing mode by pressing the "Edit" button; the details are described in the Dataset Construction section.



(a) Canvas and Function Area instructions.

(b) The two parts of sub-windows.

Figure 3.3: Combined view of Canvas, Function Area, and sub-windows.

Users can draw on the canvas and initially enable local guidance by checking the "Local" checkbox. This guidance continuously updates to match the user's drawing (it can be turned off using the "Stop" checkbox) and provides shadow hints synthesized from three retrieved sketches. For additional creative inspiration, users can click the "Generate" button, which prompts the system to provide six reference sketches (three generated and three retrieved) in sub-windows on the right side of the canvas. Users can then enable global guidance by checking the "Global" checkbox, which overlays the selected reference sketch as a red shadow, merging it with the local guidance's black shadow. This combined guidance helps users complete their drawings more effectively.

3.3.2 Dataset construction

I developed a specialized interface for collecting data on anime characters' facial features, as illustrated in Figure 3.4. This interface stored strokes as vector information and efficiently categorized stroke data according to distinct facial features. I enlisted five expert designers to create 120 freehand sketches of facial features for two genders and three age groups (six different

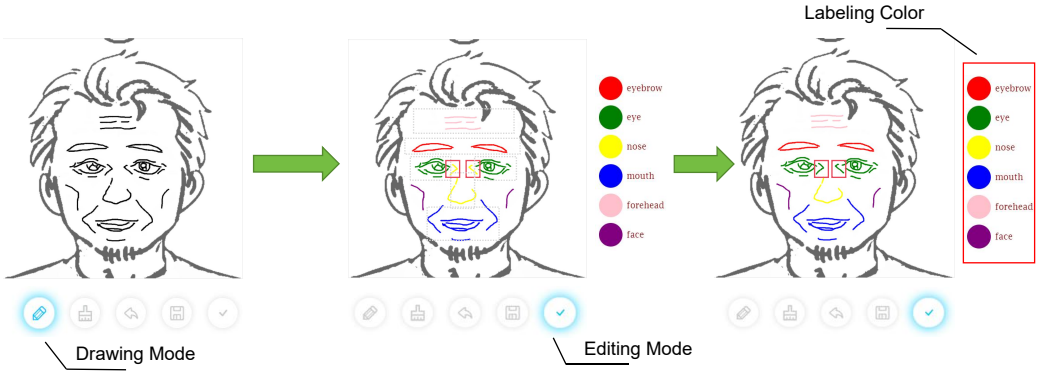


Figure 3.4: Stroke classification according to facial features: (a) the input face sketch, (b) the result of automatic stroke classification by region (gray dashed rectangle), with different colors representing different features, and (c) the final result after manual modification by the user. Strokes within the red rectangle have been corrected from incorrectly classified colors to the appropriate colors.

categories in total).

The data collection process occurred in two stages. During the drawing stage, designers completed facial feature sketches for specific gender/age characters using graphics tablets. After completing the sketch, designers moved on to the stroke editing stage, in which they classified the strokes of the drawn sketches with the assistance of the system's editing mode, as shown in Figure 3.4. The system automatically sorted all strokes by predefined regions for each facial feature. I established six facial features (eyebrows, eyes, nose, mouth, forehead, and face) and divided the corresponding regions on the canvas (the gray dashed areas in Figure 3.4). $Label(s)$ represents the automatic classification function for discriminating the label of stroke s, determined by the majority vote algorithm for each dot $d \in s$. As calculated by Eq. (3.1), the function $C_{d \in s}$ accumulates the number of dots in each stroke shared by the same facial feature. I employed the ray casting function $R(d, F)$ [101] to ascertain the facial feature region in which a single dot d is situated. F denotes the set of predefined rectangular areas in facial features.

$$label(\mathbf{s}) = \mathbf{argmax} C_{d \in \mathbf{s}}(R(d, F)) \tag{3.1}$$

Designers were tasked with identifying miscategorized strokes and manually modifying them. Ultimately, the edited data were saved to our database. Consequently, I constructed a freehand sketch database based on the strokes

of various facial features. Sample sketches from the dataset are shown in Figure 3.5.

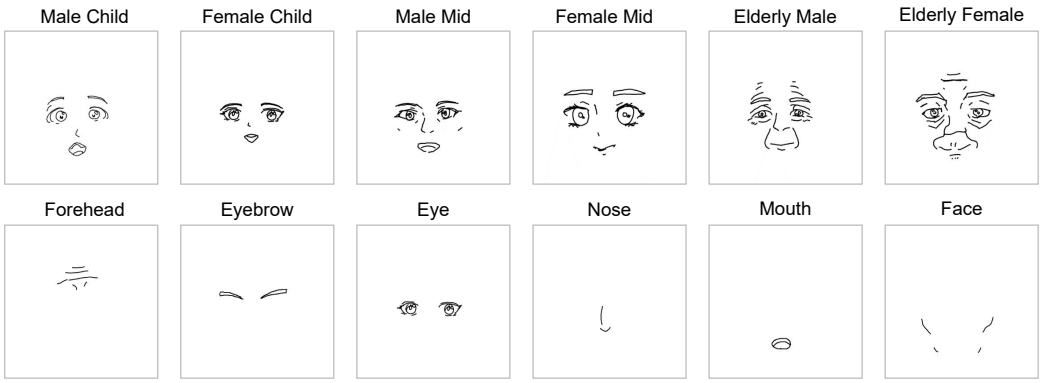


Figure 3.5: Examples of sketches in the database: The first row illustrates sketches including all facial features for global guidance, while the second row illustrates sketches of single facial features for local guidance.

3.3.3 Generation model for sketches

First, I randomly combined the six different facial features from the same age/gender category in our database, resulting in the synthesis of over 100,000 facial sketch training samples. I adopted the stroke-3 format, as proposed by Quick Draw, which represents a stroke sequence using x and y coordinates and drawing states (boolean values).

Then, I preprocessed the stroke-based sketch data, converting it into a format suitable for input into the Transformer model. This involved normalizing the stroke coordinates and encoding the sketches as sequences of tokens. To accommodate the variable-length nature of the stroke sequences, I employed positional encoding to inject positional information into the input tokens.

The model aims to combine the strengths of Transformer architecture and convolutional neural networks, leveraging the Transformer's ability to handle long-range dependencies and inherent parallelism, along with the convolutional network's capability to extract local image features. I concatenate the embedded tokens with image features extracted by a ResNet-18 network to create conditional vectors as input for the Transformer encoder, achieving feature fusion and ensuring that the generated output remains similar to the input image.

The model architecture consists of a Transformer encoder, as shown in Figure 3.6, with multiple self-attention layers and a multi-layer perceptron (MLP) decoder. I optimized the model's performance in sketch generation by increasing the number of layers and attention heads. To manage memory consumption, I employed ReZero Transformer (RZTX) layers [102]. Finally, a multi-layer MLP decoder is used to obtain the predicted tokens.

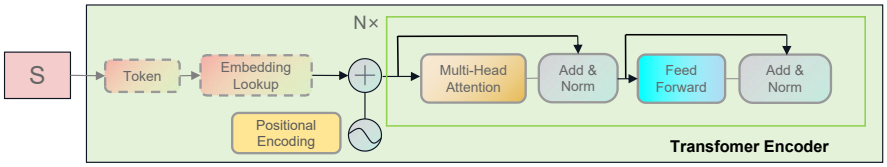


Figure 3.6: Architecture of Transformer encoder; S denotes the stroke sequences in sketch.

During the training process, I jointly trained the Transformer-based model and the ResNet-18 network, updating the model parameters concurrently. I utilized cross-entropy loss as the optimization objective for the Transformer-based model and employed a low-weight MSE loss as the objective function for the ResNet-18 network to constrain the similarity between the generated facial sketch and the input. The overall objective function is shown in Eq. (3.2):

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{MSE}}, \tag{3.2}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, and $\mathcal{L}_{\text{MSE}}$ is the mean squared error (MSE) loss between the input and output images of the ResNet-18 network. λ is a constant weight; based on experience, I set it to 0.5 during training. The total loss, $\mathcal{L}$, is the sum of the two components, each weighted by their respective coefficients.

During the inference process, I used nucleus sampling, also known as top-p sampling, to sample from the predicted logit sequence. This method adaptively selects a subset of tokens based on cumulative probability, balancing diversity and coherence. For our task, nucleus sampling ensures sketch quality while providing more diverse strokes. Extensive experiments have shown that setting the cumulative probability threshold to 0.7 achieves the best balance between generation quality and robustness.

In summary, I developed a Transformer model specifically designed for sketch generation and trained it on a pen-position sequence dataset. This approach enabled us to leverage the unique properties of the Transformer

architecture and the rich information contained in the collected data to create high-quality, diverse sketches while maintaining the required spatial similarity with the input sketches.

3.3.4 Drawing guidance

I integrated shadow guidance into our drawing interface. As the user draws a stroke on the canvas, the system retrieves or generates reference images and creates a shadow by blending the candidate sketches. Initially, I employed the Sketch-Based Image Retrieval (SBIR) algorithm proposed by Eitz et al. [103] to obtain images with contours most similar to the input line drawings. Subsequently, I utilized the shadow generation method of ShadowDraw. Key points from the drawing strokes and candidate images were extracted, and individual weights were assigned to each pixel according to the similarity between key points. By multiplying these weights by pixel values, the similarity between the image and the stroke could be expressed through pixel grayscale levels, with darker pixels denoting a higher degree of similarity to the stroke. Lastly, I overlapped the grayscale values of multiple images to produce a guidance shadow image blended from reference sketches.

As demonstrated in Figure 3.7, I combined two types of shadow guidance: global guidance, which displays all facial features, and local guidance, which shows the currently drawn facial feature. These were represented in red and black, respectively. For local guidance, I obtain references by retrieving sketches from the dataset. For global guidance, in addition to retrieval, I generate sketches to increase the diversity of reference sketches.

According to our assumptions, local guidance can reorganize different facial features, preventing premature interference during the creative process. Adding global guidance can further enhance the user's creativity by inspiring ideas for subsequent facial features. To implement local guidance, I needed to determine the category of facial features to which the current stroke belongs. This required designing a facial feature classification algorithm.

I divided the collected 540 labeled facial part stroke sequences into a training set and a validation set, with 440 strokes for training and 100 strokes for validation. Then, I processed all stroke data for the same facial feature as a single vector V using Eq. (3.3):

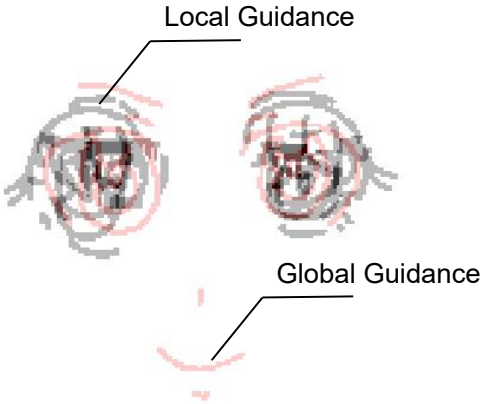


Figure 3.7: Shadow image created for global guidance and local guidance. The gray shadow is local guidance; the red shadow is global guidance.

$$V(S) = \frac{1}{K}\sum_{i=1}^{K} D_2(s_i), \text{and}$$
$$D_2(s) = \frac{1}{N}\sum_{j=1}^{N} \|d_j x + d_j y\|_2, \quad s \in S, \tag{3.3}$$

where S represents all strokes of one facial feature sketch, and s represents each stroke in S, with $S = \{s_i \mid i = 1, 2, ..., K\}$. The term $\|d_j x + d_j y\|_2$ denotes the Euclidean distance from the point in the stroke to the origin of coordinates O, where $d_j x$ and $d_j y$ represent the x-coordinate and y-coordinate values of the dots. N denotes the number of dots in one stroke, and K represents the number of strokes in a facial feature sketch.

I then employed a supervised learning algorithm, the support vector machine (SVM), to predict the facial feature classification of the strokes. The validation results demonstrated that this classification method is reliable for categorizing facial features, with an accuracy rate of 96%.

3.4 Experiment

In this section, I evaluated the effectiveness of our proposed guidance method through user studies. Additionally, I designed comparative experiments to verify the performance of the generative model, ensuring its suitability for sketch generation tasks.

3.4.1 User Study

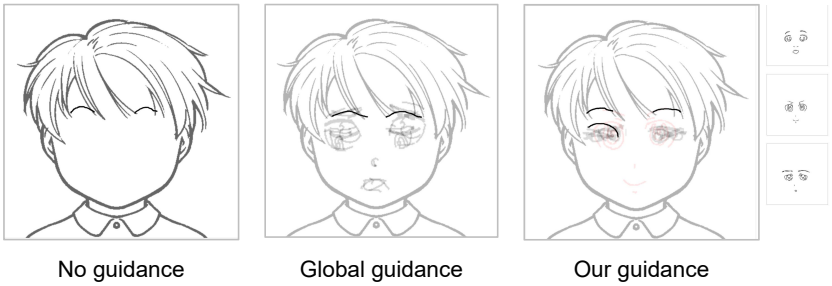
Table 3.1: User Study Questions and Results

	System Usability Scale (SUS) Questions	Mean	SD
Q1	I think that I would like to use this system frequently.	4.2	0.98
Q2	I found the system unnecessarily complex.	1.7	0.85
Q3	I thought the system was easy to use.	4.2	0.91
Q4	I think that I would need the support of a technical person to be able to use this system.	1.9	1.45
Q5	I found the various functions in this system to be well integrated.	4.4	0.73
Q6	I thought there was too much inconsistency in this system.	2.1	1.06
Q7	I would imagine that most people would learn to use this system very quickly.	4.3	0.87
Q8	I found the system very cumbersome to use.	1.9	0.68
Q9	I felt very confident using the system.	4.2	0.83
Q10	I need to learn a lot of things before I can get going with this system.	1.3	0.60
	System Evaluation Questions	Mean	SD
Q1	I think it is useful for improving my drawing skills.	4.2	0.77
Q2	I think the guidance helps draw facial features of different ages.	4.2	0.56
Q3	I think this system is helpful for adding details of facial features.	4.3	0.59
Q4	Compared to global guidance only, I think that combining global and local guidance interferes less with my creative intentions.	4.1	0.83
Q5	I think the Generate function is useful for the creation process.	4.4	1.50
Q6	I think the generated references inspire my creation process.	4.8	1.87
Q7	I think the quality of generated references is acceptable.	3.9	0.96
Q8	Compared to retrieved results, generated results better align with my design intentions.	4.1	1.53

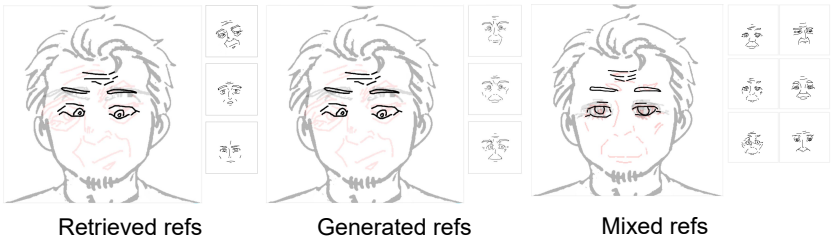
I invited 15 participants to join our user study. Based on their self-assessments, 6 participants reported high-level drawing skills, while 9 participants identified as having low-level drawing skills. All participants were required to draw portraits using Surface Pro 7. Before the user study, I explained how to use AgeFace with a user manual and allowed participants to try some warm-up exercises to familiarize themselves with the interface. Each participant was required to complete two sets of experiments: the guidance strategy evaluation experiment and the generative guidance evaluation experiment. Each set of experiments included three comparative tests. During the same set of experiments, participants had to maintain the same creative intent.

In the guidance strategy evaluation experiment, participants used three different UIs to draw facial features: one that employed our proposed guidance strategy integrating both local and global guidance, another that only incorporated global guidance to indicate the guiding methods of previous research, and a third with no guidance, as shown in Figure 3.8(a).

In the generative guidance evaluation experiment, the UI that participants used included a drawing board integrated with local and global guidance, along with right-side sub-windows displaying reference sketches. I designed three experiments to evaluate the impact of generative sketches by adding the following to the right side of the drawing board: three retrieved



(a)



(b)

Figure 3.8: Interfaces provided for guidance strategy and generative guidance evaluation experiments.

references, three generative references, and a mixed set of three retrieved and three generative sketches, as shown in Figure 3.8(b).

After completing the sketches, the participants were asked to complete a system evaluation questionnaire and the System Usability Scale (SUS) questionnaire. The system evaluation questionnaire aimed to confirm the effectiveness of the guidance strategy and the generative system, while the SUS questionnaire assessed the system's usability. The questions on both questionnaires are listed in Table 3.1. Most of the questions are rated using a 5-point Likert scale, with only two questions in the system evaluation questionnaire being single-choice, which ask about usage preferences.

Finally, I invited three evaluators who did not participate in the user study (one had high-level drawing skills, and the other two had low-level drawing skills) to conduct subjective ratings of all sketches (a total of 90 sketches) based on two aspects: sketch quality and sketch similarity. For sketch quality, I asked the evaluators to rate each sketch using a 5-point Likert scale (1 = "very poor"; 5 = "very good"). For sketch similarity, I set up four groups of sketch comparisons across the two experiments: no guidance vs. global guidance, no guidance vs. local-global guidance, generative references vs. mixed references, and retrieved references vs. mixed references. The evaluators rated the similarity of facial features in each comparison group

using a 5-point Likert scale (1 = "completely different"; 5 = "exactly the same"). The results of these two evaluations are shown in Figures 3.9 and 3.10.

I also illustrate some representative results of the user study. The first set of results, illustrating the effectiveness of the guidance strategy, is shown in Figure 3.11. The second set of results, validating the effectiveness of the generative model, is shown in Figure 3.12.

3.4.2 Comparative Experiments

I conducted comparative experiments to validate the superiority of our generative model against three classic baselines: SketchRNN [100], Sketchformer [31], and Pix2Pix [104]. SketchRNN, a sequential generation model based on an RNN architecture, uses the same input format as our model. Sketchformer is another sequential generation model based on the Transformer architecture. Pix2Pix is a classic conditional image generation model based on GANs. I validated the superiority of our model by conducting a quantitative evaluation, comparing the diversity and similarity of sketches generated by the three different models to the input sketches.

To ensure experimental fairness, all models were preprocessed and trained using the same dataset. Specifically, since Pix2Pix is a pixel-level generative model, I preprocessed the data by masking parts of the stroke sequences and converting them into pixel images. These paired images, consisting of masked and complete sketches, formed the training set.

All training was performed on a single NVIDIA 3090 GPU. The objective function's hyperparameter λ was set to 0.1, and the maximum token length was configured to 512. Under these settings, the models achieved an optimal balance between robustness and generation quality. I extracted two sketches from untrained data and removed some strokes, as shown in Figure 3.13. These modified sketches were used as inputs to generate 100 sketches each with three different models (a total of 300 sketches), forming a validation set. For SketchRNN and Sketchformer, which do not require an input sketch, I used the category labels "child&male" and "elderly&male" as conditions for generation. For Pix2Pix, I converted the stroke sequences into pixel images to use as inputs for generation.

I utilized a traditional perceptual hashing algorithm to calculate the pairwise distances (PD) between sketches to evaluate the diversity of the generated results. A larger average distance indicates greater diversity. On the other hand, I used the F1 score as a metric to evaluate sketch similarity. This metric is well-suited for evaluating the similarity of binary data. Due to the sparse matrix nature of sketches, with a large amount of white

background causing minimal result differences, background pixels were set to 0, and foreground pixels were set to 1. This way, the white background was treated as non-informative, focusing only on the black sketch parts of the image. By calculating the F1 score of these pixels, I assessed the similarity between the input sketches and the generated sketches.

3.4.3 Result Analysis

3.4.3.1 User Study Analysis

I analyzed the results of the two user studies described above. In the guidance strategy evaluation experiment, subjective rating scores for sketch quality are shown on the left side of Figure 3.9. The results indicate that using our proposed global and local guidance strategy significantly improved the sketch quality for all participants compared to no guidance and global guidance only. The subjective ratings for sketch similarity are shown on the left side of Figure 3.10. The results demonstrate that sketches created with global and local guidance were more consistent with the initial drawing intent (facial expressions and features) than those without guidance. In contrast, global guidance only interfered with the creative process. This confirms our hypothesis that adding local guidance better maintains the user's creative intent. Lastly, by using the local guidance strategy, features from different sketches were fused to generate more diverse character images.

In the generative guidance evaluation experiment, the subjective ratings for sketch quality are shown on the right side of Figure 3.9. The results indicate that sketches drawn using a mix of generative and retrieved references were of higher quality than those with a single reference type. Although sketches drawn with generative references were of slightly lower quality compared to those with retrieved references, the subjective ratings for sketch similarity (right side of Figure 3.10) show that the similarity of sketches drawn with mixed and generative references was higher than those drawn with retrieved references. This suggests that users preferred using generative references, as they provided better support for their creative intent.

The qualitative analysis of user sketches also yielded similar results. In Figure 3.11, users achieved more detailed sketches with drawing guidance compared to those with no guidance; furthermore, they retained the original intent more effectively when using local-global guidance compared to using global guidance only. Figure 3.12 illustrates how sketches created with generative guidance closely resembled those created with mixed references.

For the results of the SUS questionnaire in Table 3.1, according to the standard SUS calculation method, the scores for odd-numbered questions

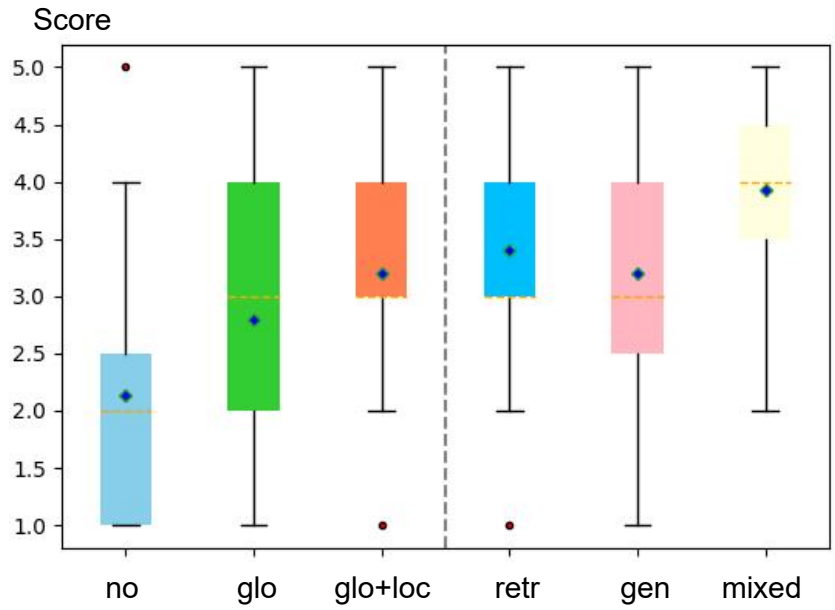


Figure 3.9: Subjective scores for sketch quality. The vertical axis represents evaluation scores, and the horizontal axis represents different experimental groups, listed from left to right as follows: No Guidance, Global Guidance Only, Global-Local Guidance, Sub-window with Retrieved References Only, Sub-window with Generated References Only, and Sub-window with Both Retrieved and Generated References.

were reduced by 1, and the scores for even-numbered questions were subtracted from 5. All converted scores were then summed and multiplied by 2.5. The system's SUS score was 81, indicating good usability. For the system evaluation questionnaire in Table 3.1, the average score for the eight rating questions was 4.25, with a median score of 4.2. This shows that most users found that incorporating generative sketches as references effectively supported their design process. In the other two single-choice questions (Which guidance do you prefer as the reference while drawing? During the drawing process, which sketches do you prefer to refer to?), three users preferred using global guidance, three users preferred using local guidance, and nine users preferred using both global and local guidance. For sketch references, three users preferred retrieved sketches, six preferred generated sketches, and six preferred using both retrieved and generated sketches equally during the drawing process. This also demonstrates that our proposed guidance strategy and generative model were the most favored during the user's creation process.

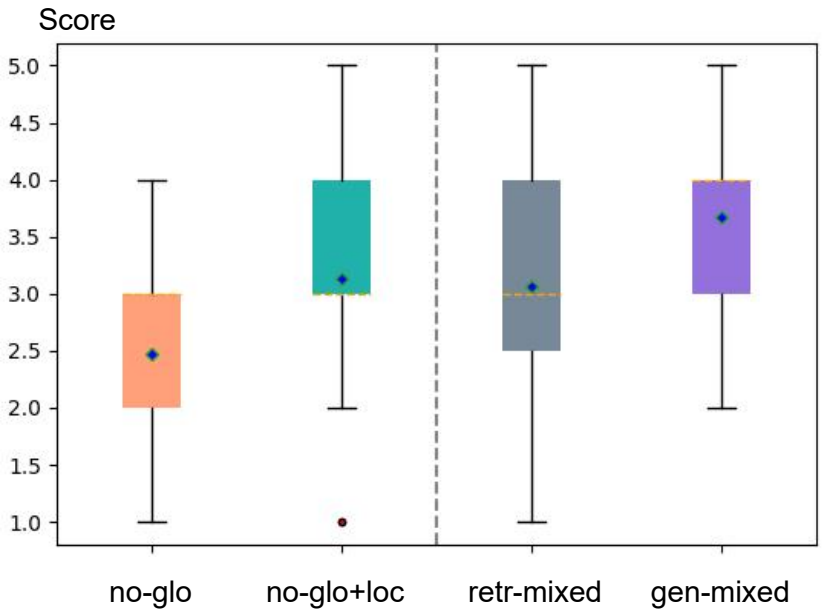


Figure 3.10: Subjective scores for sketch similarity. The vertical axis represents similarity scores between sketches, and the horizontal axis represents comparisons of different experimental groups, listed from left to right as follows: No Guidance vs. Global Guidance Only, No Guidance vs. Global-Local Guidance, Retrieved References Only vs. Both Retrieved and Generated References, and Generated References Only vs. Both Retrieved and Generated References.

3.4.3.2 Comparative Experimental Analysis

All models were trained and evaluated on an identical dataset comprising multi-age anime face sketches, ensuring consistency and fairness in the comparison. The models were classified into two categories: sequence-to-sequence and pixel-level models, with each type undergoing supervised training using paired data. For sequence-to-sequence models, partial sequences were masked and provided as inputs, while the complete sequences served as the ground truth. Conversely, pixel-level models received partially occluded images as inputs and were trained to generate the full images as outputs. To maintain equitable evaluation across different model architectures, all validation metrics were based on pixel-level calculations, ensuring that the performance assessments were fair and comparable across both model types.

The experimental results are shown in Table 3.2. By comparing pairwise distances, I found that SketchRNN produced the highest diversity in its generated results. However, as illustrated in Figure 3.13, these results often

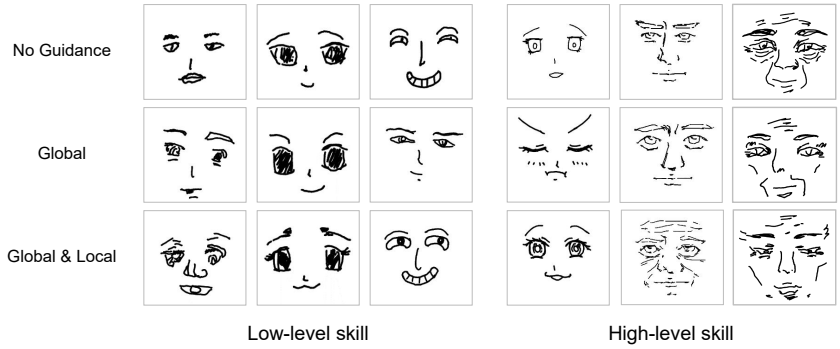


Figure 3.11: Drawing results from participants with different drawing skills. Each column displays sketches from the same participant.

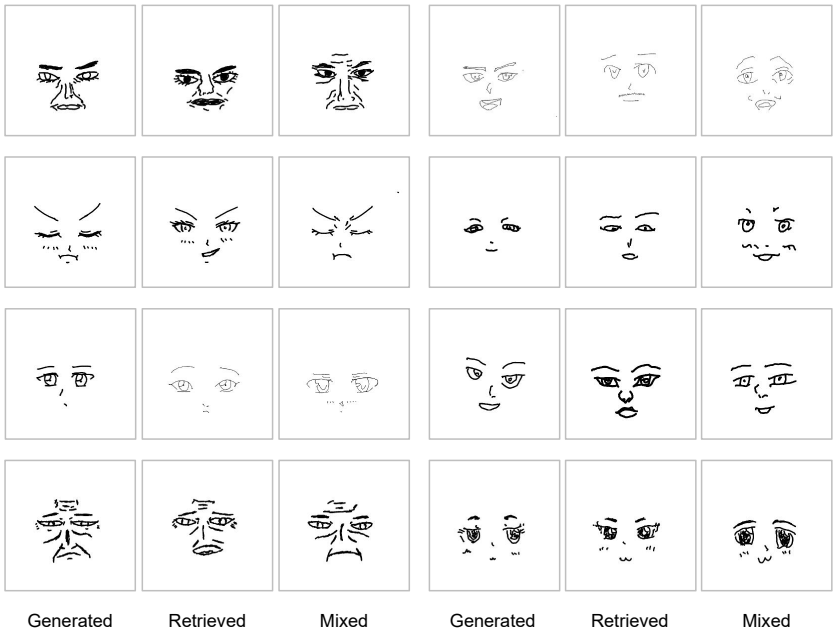


Figure 3.12: The second set of user study results showcasing participants' drawings. The left column displays sketches created with generated references. The middle column shows sketches guided by retrieved references. The right column presents sketches drawn using a mix of generative and retrieved references.

Figure 3.13: Generated results from the four models compared with the input sketches under the categories of male child and elderly male; for the sequential generation model, I used different colors to represent different strokes.

suffered from poor quality and failed to ensure a reasonable spatial stroke structure. Sketchformer's generation quality was better, but it faced similar issues as SketchRNN. In contrast, our model achieved a good balance between quality and diversity.

According to the F1 score, Pix2Pix's generated results showed the highest similarity to the input sketches. However, its low pairwise distances indicated a lack of diversity in the generated sketches, and the overall quality of its generated images was also poor.

The experimental results demonstrate that our model achieved the best balance among generation quality, diversity, and similarity to the input. This makes it the most suitable generative model for drawing guidance.

Table 3.2: Comparative Experimental Results

Model	PD ↑	F1 Score ↑
SketchRNN	**27.11**	0.05
Sketchformer	23.21	0.13
AgeFace	17.86	0.24
Pix2Pix	10.3	**0.60**

3.4.3.3 Additional Experimental Results

In this section, I provide further details on our model design.

Our experiments showed that deeper ResNet architectures like ResNet50 and ResNet101 did not significantly improve performance. Therefore, I chose the more efficient ResNet18, which reduces parameters while maintaining accuracy, serving as the backbone for feature extraction to ensure streamlined learning and faster processing. In addition to the ResNet backbone, I incorporated the RZTX transformer layer into our model's architecture. The RZTX layer, known for its lightweight design and improved convergence properties, replaces the standard transformer layers typically used in vision transformer models. As illustrated in Figure 3.14, the combination of ResNet18 and the RZTX transformer layer results in faster convergence compared to models utilizing deeper ResNet architectures or standard transformer layers. This design choice allows the model to achieve better training efficiency while maintaining strong performance in various tasks.

Figure 3.14: The loss variation over the first 2000 iterations when using the RZTX layer and the standard transformer layer.

3.5 Conclusion

In this paper, I propose AgeFace, a drawing support system to help users sketch facial features for multiple ages. I constructed an age-related facial feature dataset based on stroke data and introduced a novel drawing guidance strategy to support users' design intents. Additionally, I trained a generative model based on a Transformer architecture on our dataset. By incorporating CNN image feature extraction techniques, I supported a broader range of anime character facial features and generated results that preserved the identity features of the input sketches. The effectiveness and advancement of our model were validated through user experiments and baseline comparison experiments.

Additionally, Designing multi-age facial features for anime characters presents a key challenge in balancing **Precision** and **Manipulability**, where users require detailed, accurate facial features while retaining creative control.

The interactive drawing interface supports real-time user adjustments through continuous input and feedback loops, enhancing system **manipulability**. Experimental evaluations highlight the effectiveness of this design: results from the System Usability Scale (SUS) and functionality questionnaires show that users found the system highly interactive and supportive of their creative process. User study assessing sketch similarity demonstrate that the generated outputs closely align with user inputs, significantly outperforming retrieval-only methods in **output precision**. Additionally, metrics such as F1 scores and visualizations confirm that the system produces facial sketches with high similarity and design fidelity. This seamless integration of retrieval-based detail precision and model-driven generative adaptability enables AgeFace to resolve the inherent contradiction between precision and manipulability, supporting both creative freedom and professional-level design quality.

3.6 Limitations and Future Works

Despite its strengths, AgeFace has limitations. The system struggles to support certain features, such as hair and glasses, due to the limited scope of the dataset. Additionally, the lack of multi-view sketches in the database hinders the generation of consistent facial sketches from different angles. Future work will focus on expanding the database with more diverse features and multi-view sketches, optimizing the model architecture to reduce inference time, and enabling real-time, responsive guidance through enhanced adaptability

to user interactions, such as zooming, rotating, and screen adjustments. These enhancements aim to further reconcile precision and manipulability, empowering users in their creative workflows.

Chapter 4

Character Pose Design

In recent years, the widespread application of conditional diffusion models in image generation has led to significant advancements in **controllable human image generation (HIG)**, including both face and whole-body generation. Despite these improvements, existing conditional diffusion models, such as ControlNet and T2I-Adapter, still face challenges of ambiguous condition inputs and insufficient conditional guidance when utilizing a single denoising loss. These limitations result in issues with condition recognition and accuracy in HIG tasks. Consequently, current methods often require extensive iterative trial-and-error processes in character design workflows, making it difficult to achieve a satisfactory balance between the key elements of **Convenience** and **Precision** in these application scenarios.

To address these challenges, I introduce two innovative solutions specifically designed for parametric signal-guided generation tasks, such as those involving pose keypoints and facial landmarks. Firstly, I propose the **Spatial Guidance Injector (SGI)**, which enhances conditional details by encoding text inputs with parametric signals. This approach provides clear, annotated guidance, thereby resolving issues associated with using only image features as ambiguous control inputs. Secondly, to overcome the limitations of conditional supervision, I introduce **Diffusion Consistency Loss (DCL)**. DCL applies supervision to the denoised latent code at any time step, promoting consistency between the latent code at each step and the input signal. This method improves the robustness and accuracy of the output. The combination of SGI and DCL forms our **Effective Controllable Network (ECNet)**, which offers more precise conditioning input and stronger controllable supervision within an end-to-end text-to-image generation framework.

Additionally, DCL, as a general method, can extend to other structural conditions, such as Canny and segmentation, and enhance the control accuracy effectively. Extensive experiments conditioned on human skeletons and facial landmarks demonstrate that ECNet significantly enhances the controllability and robustness of generated images.

Figure 4.1: The core work of this paper is to design a general framework for supervised training of diffusion models, and enhancing the controllability of text-to-image diffusion models. The figures show three categories conditions: skeleton, facial landmark, and canny. Each category includes: (I) the original image used for reference; (II) the conditional image derived from the original image; (III) results generated by ControlNet; (IV) comparison results generated by the state of the art model, HumanSD(skeleton and landmarks) and ControlNet++(canny); (V) results generated by ECNet (our model). Compared to other SD-based models, our model ECNet exhibits superior capabilities and robustness in image generation with control across all categories. In Canny Edge results, the areas within the orange boxes highlight regions with low control precision.

4.1 Introduction

Controllable Image Generation is a critical area of research in computer vision and deep learning, with significant recent advancements [52, 54, 105–108]. The capacity to synthesize images that conform to predetermined conditions not only extends the frontiers of conventional image synthesis methodologies but also serves an array of application-specific demands. Such technological advancements are of paramount importance in disciplines such as virtual reality, film production, and fashion design, where automating image creation tailored to particular themes can substantially augment efficiency and mitigate expenses.

Due to the unparalleled performance of diffusion models in text-to-image generation, they have outperformed the results generated by Genera-

tive Adversarial Networks (GANs) [109–111] and Variational Autoencoders (VAEs) [112,113] in image generation. The forefront of controllable diffusion models, epitomized by ControlNet [52] and T2I-Adapter [54], has realized a measure of control in image generation. These models incorporate various constraints, throughout the generation process, significantly improving the controllability of the base SD model. Recent advancements like HumanSD [105] and Composer [114] further refine this approach by integrating additional conditions into the noisy latent embeddings used by the SD U-Net module. This leads to more stable training and enhanced model robustness

Despite these improvements, current diffusion models still face challenges. For instance, the leading skeleton-guided diffusion model, HumanSD, struggles with generating images in intricate scenes, dynamic actions, and nuanced details. This limitation primarily stems from the limited controllability inherent in the end-to-end training methodology, which includes issues such as the ambiguity of condition inputs and the lack of comprehensive conditional supervision beyond a singular denoising loss. Our contemporaneous work, ControlNet++ [115], introduces an approach that explicitly optimizes pixel-level cycle consistency between the denoised model output and the ground truth. However, due to limitations in the denoising method, it is restricted to a smaller range of time steps (timestep<200), making it difficult to achieve consistent supervision throughout the range of time steps. To resolve the main issues of the state-of-the-art models, I introduce two innovative solutions.

First, the Spatial Guidance Injector (SGI) incorporates precise annotation information as a condition, complementing the condition image, as shown in Figure 4.2. Annotations serve as an effective means to define human postures and facial orientations, providing detailed context that, when combined with the global structure from the condition image, offers a comprehensive understanding of these features. Our approach integrates image conditions with annotation and text conditions, where the image is processed through a U-Net for global feature extraction, while annotations and text are combined via the *SGI* architecture to enrich contextual detail.

Second, I propose Diffusion Consistency Loss (DCL), which uses the denoised latent code for loss calculation instead of focusing solely on noise, as illustrated in Figure 4.2. This approach provides more accurate guidance by comparing the model output to a closer approximation of the ground truth. Existing methods struggle to maintain high fidelity across all diffusion process time steps, as shown in Figure 4.3, making it difficult to apply consistent supervision. Our *DCL* introduces a dual-stage loss formulation, adaptable throughout the denoising process, enhancing supervision and contributing to a more stable training process.

The main contributions of this paper are:

- I proposed ECNet, an innovative framework for controllable human image generation. I first identified that parametric signals provide a clearer condition for HGI compared to the ambiguous manipulation of image features alone. I introduced a Spatial Guidance Injector (SGI) architecture, enhancing input control and improving contextual depth and image controllability.
- The development of a novel Diffusion Consistency Loss (DCL) within ECNet. *DCL* is the first to utilize denoised latent code for supervision and incorporates dual loss formulations tailored for different stages of the training process. This significantly boosts the model's controllability and robustness of its outputs.
- The efficacy and efficiency of the ECNet framework are validated through various evaluation metrics across multiple domains, including skeletons, landmarks, sketches, and Canny, shown in Figure 4.1. The performance of ECNet surpasses previous state-of-the-art models in a fair experimental setting.

4.2 Related Works

4.2.1 Text-to-Image Diffusion Model

Diffusion models have established themselves as state-of-the-art in deep generative modeling, outperforming Generative Adversarial Networks (GANs) [116] in image synthesis tasks. Benefiting from the remarkable ability of large-scale language models, such as CLIP [27], to encode textual inputs into latent vectors, diffusion models have demonstrated astonishing capabilities in text-to-image generation tasks. For instance, one of the earliest text-to-image diffusion models, Glide [117], is a text-guided diffusion model that also supports image super-resolution generation and editing. Imagen [42], a text-to-image architecture, discovered significant improvements using a pre-trained large-scale text-only encoder and introduced a new Efficient U-Net structure. Latent Diffusion Model(LDM) [118] was the first to propose conducting diffusion and reverse diffusion in feature space, significantly enhancing efficiency, and introduced the use of cross-attention to embed conditional information, allowing for more flexible incorporation of conditions. Stable Diffusion, a large-scale implementation of latent diffusion, was developed for text-to-image tasks. However, all these models typically only take text as input, making it challenging for precise image control, such

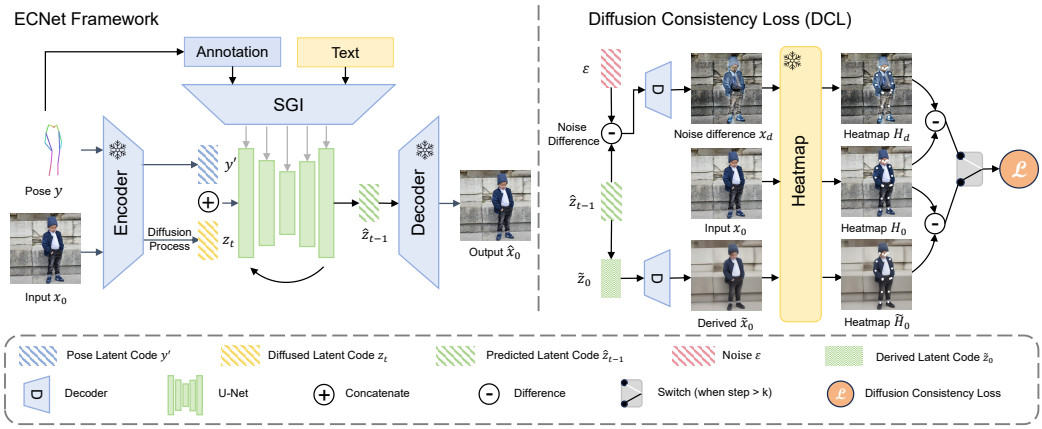


Figure 4.2: The framework and its loss design are illustrated using the task of skeleton control as an example. our model encodes the skeleton image into a latent code via a VAE to obtain a pose latent code. This code combines with diffusion's noise code as input for a U-Net. Additionally, Our *SGI* module further combines corresponding pose annotations and text, integrating them into the U-Net layers. During the training phase, I enhance the conditional generation capabilities of the diffusion model by introducing *DCL*. *DCL* targets heatmap disparities between estimated and input images, using dual-stage loss to impose consistency supervision throughout the diffusion process. z represents the latent code and x denotes the image decoded from z. Please refer to 4.4 for more details

as target positioning and posture control. Consequently, subsequent works have seen the emergence of numerous studies focusing on controlled image generation using diffusion models.

4.2.2 Controllable Diffusion Model Generation

In this section, I primarily focus on controllable diffusion model generation, specifically on how to incorporate additional conditions into text-to-image models, such as bounding boxes, human poses, and sketches. Among the most influential works in this area are ControlNet [52] and T2I Adapter [54]. Both of them fix the original weights of Stable Diffusion and use an additional trained branch to modify the embeddings in the U-Net for guiding generation. I refer to this approach as the dual-branch diffusion model. This method supports a variety of conditions, including human poses, Canny Edge Maps, and more, enabling flexible image generation control.

Additionally, there has been a surge of recent works targeting different tasks. Uni-ControlNet [119] and Composer [114] address image generation

control under multiple conditions, considering the interrelations between different conditions. They categorize conditions into local and global, adopting different methods of integration depending on the type of condition. Both LayoutDiffusion [120] and GLIGEN [121] use bounding boxes (bbox) as conditions for controllable generation. LayoutDiffusion integrates encoded bbox information into the U-Net, merging image and layout features for controllable generation. In contrast, GLIGEN adds new attention layers to handle the fusion of bbox and text without altering the original weights of Stable Diffusion, endowing the model with the capability to control generation using bbox.

4.2.3 Conditional Human Image Genteration

Conditional HIG involves creating realistic human images conditioned on specified body or facial guidance images. Traditional approaches utilized GANs [122, 123] and Variational Autoencoders(VAEs) [124, 125]. These models offered foundational advancements but often struggled with maintaining fine-grained details and variability in condition adherence. Diffusion models have emerged as a powerful alternative in image synthesis due to their inherent capabilities for producing detailed and high-resolution images. Controllable diffusion models are now widely applied in HIG [105, 126–128], achieving significant improvements in image quality and fidelity compared to previous methods.

HumanDiffusion [126] introduced a coarse-to-fine alignment diffusion framework to enhance the alignment quality from the image level to the feature level and from low to high resolution. This approach achieves good performance even in complex tasks with diverse details and uncommon poses. However, the model relies on high-quality pose pairing as a condition, making it difficult to apply to in-the-wild datasets. HumanSD [105] proposed an alternative approach for adding conditions. This work combined the pose image embedding with the noisy image embedding as the input to the U-Net for training, showing superior pose control capabilities compared to dual-branch diffusion models. Additionally, it also optimized the original SD loss by incorporating a weight more focused on human posture to generate results more aligned with the pose conditions. Although the loss was optimized, this method essentially still achieves control by adjusting the input method.

Current pose-guidance frameworks like ControlNet, HumanSD, and sketch-guided diffusion models frequently encounter challenges in generation accuracy, especially when interpreting complex spatial prompts or multi-character scenarios. This limitation reflects a struggle to achieve Precision while retaining the Convenience for existing methods.

Text-guided models like Stable Diffusion and DALLE-2 provide a high level of convenience by using simple text prompts, yet they often fall short in generating spatially accurate poses. Their reliance on textual descriptions alone can lead to mismatches between the intended and generated poses due to limited pose-specific representation learning.

Our method resolves this Precision-Convenience contradiction through two key innovations: the Spatial Guidance Injector (SGI), which efficiently integrates structural guidance into the diffusion process, and the Diffusion Consistency Loss (DCL), which ensures precise pose generation by enforcing powerful supervision. This framework allows efficient and accurate pose generation without additional input requirements, offering a well-balanced between precision and convenience.

4.3 Preliminaries and Motivation

This section discusses the issues of existing methods and the inspiration of ECNet. These methods uniformly adopt the Latent Diffusion Model (LDM) as their foundational framework, capitalizing on its high trainability and exceptional generative quality, while employing various control schemes. These methods are introduced in Section 4.3.1. Subsequently, Section 4.3.2 elucidates the problems present in these methods and the motivation behind designing ECNet.

4.3.1 Preliminary Introduction

The training process of the Diffusion model is conceptualized as a standard diffusion process, where an input latent code z_0 incrementally acquires noise over t time steps, transitioning into a latent code close z_t approximating random noise. This process is mathematically articulated as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{4.1}$$

where $\bar{\alpha}_t$ denotes a predetermined noise level coefficient, ϵ represents noise drawn from a standard normal distribution, and t signifies the time step.

During the denoising phase, the model learns to predict the input latent code z_0 from diffused latent code z_t, a process achieved by optimizing the following objective function:

$$\mathcal{L} = \mathop{\mathbb{E}}_{t,z,\epsilon} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2 \right], \tag{4.2}$$

Where ϵ_θ is the noise predicted by the model, to minimize the discrepancy between the predicted noise and the actual noise. Through this mechanism,

the Stable Diffusion Model effectively learns the data distribution and generates high-quality images.

Inspired by previous work [129], a novel method controls diffusion model inference via sketches, differing from typical condition-based control. It utilizes a pre-trained edge predictor during training for mapping noisy image features to edge maps. At each denoising step t, features are input into a latent edge predictor to estimate edge maps, with the similarity gradient between predicted and true edges guiding denoising, as shown in Equation 4.3. This edge guidance ensures synthesized images closely align with the target edges.

$$\hat{z}'_{t-1} = \hat{z}_{t-1} + k \cdot \nabla_{z_t} Loss \tag{4.3}$$

where $\hat{z}_{t-1}$ represents predicted latent code at time step $t-1$, $\hat{z}'_{t-1}$ denotes the predicted latent code after guided by the gradient, $Loss$ denotes the calculated edge loss, and k governs the intensity of the guidance exerted by the loss.

4.3.2 Motivation

Existing SD-based control models predominantly use condition images as their primary input for control. However, relying entirely on image features for control conditions does not provide adequate guidance, leading to certain details in the generated images being controlled imprecisely. In contrast to images, annotations include sequences and coordinates, complementing the global control provided by textual conditions, and giving a more detailed guide for image generation. Therefore, I suggest that incorporating image annotation information into traditional textual conditions can significantly improve the controllability of the generated outcomes.

Beyond the issue of insufficient conditional inputs, existing diffusion models also suffer from a lack of supervision on conditions or from employing ineffective supervisory methods. For instance, ControlNet uses traditional stable diffusion loss without any condition-based supervision. While HumanSD uses a heatmap-guided weighted loss to strengthen the new structure-aware condition by weighting the original diffusion loss with an estimated noise difference heatmap. However, it still applies to the supervision of noise without direct supervision over the latent code. To develop supervision on the denoised latent code at any time steps of the diffusion model, I explore a novel method to estimate the image during the denoising process.

The diffusion process is typically fixed and employs a predefined variance schedule, allowing for the sampling of z_t at any time step t directly from

z_0, as illustrated by Equation 4.1. Consequently, I can deduce the result as shown in Equation 4.4.

$$\tilde{x}_0 = \frac{\hat{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\varepsilon}}{\sqrt{\bar{\alpha}_t}} \tag{4.4}$$

$$\tilde{x}_0 = \frac{\hat{\varepsilon} - \sqrt{1-\bar{\alpha}_t}\varepsilon}{\sqrt{\bar{\alpha}_t}} \tag{4.5}$$

Equation 4.4 describes the derivation process for the denoised image, while Equation 4.5 describes the noise difference image. In both equations: $\tilde{x}_0$ represents the derived initial sample, $\hat{\varepsilon}$ denotes the model-predicted noise, ε refers to Gaussian noise. Note that in our paper, z is used to represent the latent code, and x denotes the image decoded from the latent code.

In Equation 4.4, I derive an approximate $\tilde{x}_0$ directly fromx_t at any timestep t. For the noise difference image in Equation 4.5, Gaussian noise replaces the model-predicted noise, and x_t is substituted with the predicted noise. At larger timesteps, $\hat{\varepsilon}$ better approximates x_t and noise level similar to Gaussian noise. Gaussian noise, allowing Equation 4.5 to align with Equation 4.4 under these conditions. Note that using idealized Gaussian noise instead of model-predicted noise at larger timesteps results in a derived $\tilde{x}_0$ from Equation 4.5 that is closer to the target than from Equation 4.4.

As shown in the third row of Figure 4.3, it is observed that global image features, such as foreground-background separation and pose structure, are primarily generated in the initial phases of the denoising process, while local features predominantly occur during the later stages.

Therefore, An intuitive approach involves supervising the heatmap features of the derived image $\tilde{x}_0$ against the input image, with supervision intensity dynamically adjusted based on time steps. Specifically, during the early time steps, the predicted image more closely resembles the original image, leading to a reduced error in heatmap detection. However, due to the lower quality of the derived images $\tilde{x}_0$ at larger time steps, significant errors occur in keypoint detection results, as shown in Figure 4.3 for images derived at steps 800 and 900, thereby diminishing the accuracy of the supervision.

To mitigate this challenge, I introduce noise difference code, as shown in the first row of Figure 4.3. The principle behind the noise difference code and derived latent code is the same, as both use Equation 4.1 to obtain the initial image from the predicted noisy image. The key difference is that the noise difference code uses random Gaussian noise instead of the noisy latent code at the current time step. This means that at larger time steps (the noisy latent code is closer to Gaussian noise), the predicted image is closer to the original image, which provides more accurate keypoint heatmaps at

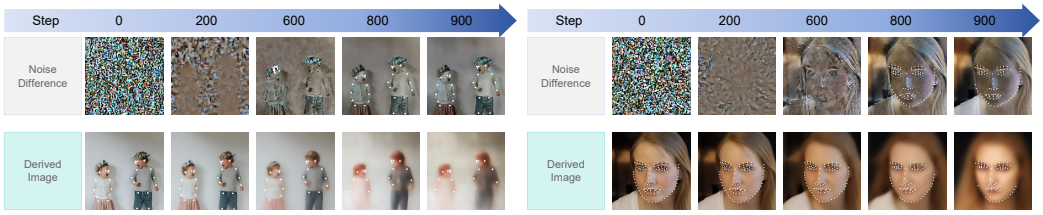


Figure 4.3: The decoded images of pose and face at different time steps. The first row shows the decoded images obtained from the noise difference code. The second row displays the denoised results derived from the predicted noise latent code. The white points indicate keypoints detected using the pre-trained detector provided by MMPosee [2].

larger time steps. Noise difference code ensures more effective and robust conditional supervision during these later stages of the denoising process.

4.4 Method

4.4.1 Diffusion Consistency Loss

Inspired by existing studies, I propose the integration of additional latent code supervision into the general loss structure applicable for classifier-free guidance, to enhance the generation accuracy. The entire loss $\mathcal{L}_e$, as depicted in Equation 4.6, is divided into two primary components, termed the original SD loss $\mathcal{L}_{SD}$ and Diffusion Consistency Loss $\mathcal{L}_{DC}$.

$$\mathcal{L}_e = \mathcal{L}_{SD} + \alpha \mathcal{L}_{DC} \tag{4.6}$$

According to Section 4.3.2, the construction of $\mathcal{L}_{DC}$ adopts distinct supervision strategies during two phases of the diffusion process, as illustrated in Equation 4.7. This loss design harnesses the high fidelity of the noise difference image and the derived image $\tilde{x}_0$ at different timesteps, providing precise supervision for the training process.

$$\begin{aligned} &\mathcal{L}_{DC} = \mathcal{L}_{drv} \text{ if } t < k, \text{ else } \mathcal{L}_{dff} \\ &\mathcal{L}_{drv} = |H_{inp} - H_{drv}|, \quad \mathcal{L}_{dff} = |H_{inp} - H_{dff}| \end{aligned} \tag{4.7}$$

Where H represents the heatmap features of images decoded from latent code. H_{inp}, H_{drv}, and H_{dff} mean the heatmap features of the input images, the derived images $\tilde{x}_0$, and the noise difference images. $\mathcal{L}_{drv}$ and $\mathcal{L}_{dff}$ are both L1 losses, $\mathcal{L}_{drv}$ is active for $t < k$, emphasizing the alignment between input image and $\tilde{x}_0$ at earlier time steps. $\mathcal{L}_{dff}$ is relevant for $t \geq k$, focusing

on mitigating the negative impact caused by keypoint detection errors in lower-quality $\tilde{x}_0$ in later time steps k. The value of k is set as different time steps for different tasks. More details are provided in the Supplementary Material, and α is a constant item to control the supervision intensity. The working principle of $\mathcal{L}_{DC}$ is the same as Equation 4.3.

Notably, our approach to the noise difference method was developed independently through mathematical derivation, resulting in a significantly different understanding and application compared to HumanSD:

1. Comparing with HumanSD performs $\hat{\varepsilon} - \varepsilon$ intuitively, I provide a mathematically rigorous explanation of its underlying principles, yielding a much clearer image quality of noise difference images.

2. Unlike HumanSD, which does not leverage this method for direct supervision of control conditions, our approach introduces the consistency loss specifically designed for supervising SD-based control models, which significantly enhances the performance of our method.

4.4.2 Spatial Guidance Injector

Previous Sd-based pose control models employed skeletal images to incorporate pose conditions, utilizing a VAE module to process these skeletal images for positional information, ensuring alignment of pose conditions with the latent embedding of input images. However, I suppose that extracting image features to derive pose information is rather indirect. In contrast, the keypoint annotation embedded within skeletal images offers more direct spatial information for pose representation. Moreover, I observed that textual conditions typically do not encompass specific details such as the number of objects or joint positions. Given those, I propose integrating the keypoint annotations as an additional condition to the existing posture image and textual conditions. Specifically, each image is processed to extract keypoint annotations, which are then padded to a shape of "batch size, the maximum number of objects, the number of keypoints, coordinate dimensions". A mask is applied to the padded positions. I then flatten the coordinates of all keypoints for each detected object and embed them into the same dimension as the text embeddings generated by the CLIP encoder. To synthesize the visual and textual information, I employ a self-attention mechanism along the dimension of the detected objects and integrate the results with the text embeddings via a cross-attention module. This integrated module is called *Spatial Guidance Injector*(SGI), as the Equation 4.8. The *SGI* facilitates a more sophisticated understanding of the multimodal annotation data.

Model	Pose Performance Metrics					Face Performance Metrics		
	AP(%)↑	CAP(%)↑	PCE↓	CLIPSIM↑	FID↓	NME↓	CLIPSIM↑	FID↓
ControlNet(SD2.1)	22.06	62.38	1.45	33.53	4.59	0.37	30.13	6.09
HumanSD(SD2.1)	33.15	63.48	1.43	32.63	4.74	0.46	29.89	3.95
ControlNet(SDXL)	20.49	64.33	1.89	**33.67**	**4.13**	0.24	**30.91**	3.75
ECNet(SD2.1)	43.31	65.76	1.35	32.28	4.89	0.33	29.46	**3.21**
ECNet(SDXL)	**46.30**	**66.70**	**1.32**	32.93	4.88	**0.20**	30.84	4.11

Table 4.1: Quantitative comparisons between ECNet and other SD-based models. I conduct experiments on ECNet for two primary tasks: human skeleton control and facial landmark control. The results indicate that ECNet outperforms previous SD-based models in both tasks.

$$\text{Softmax}\left(\frac{\mathbf{W}_Q C(t)(\mathbf{W}_K A(a))^\top}{\sqrt{d_k}}\right)(\mathbf{W}_V A(a)) + C(t) \tag{4.8}$$

Where, $A(a)$ symbolizes the self-attention mechanism applied to the annotations. $C(t)$ denotes the frozen CLIP encoder that extracts meaningful textual features from prompts. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the weight matrices for the Query, Key, and Value in the attention mechanism, respectively. These matrices transform the inputs into representations suitable for generating attention scores, and d_k is the dimension of the key vectors. The softmax function applied to the cross-attention result between $A(a)$ and $C(t)$, produces a distribution representing the attention weights. The final output is derived by multiplying the result of the softmax function with the value matrix $(\mathbf{W}_V A(a))$ and adding it to $C(t)$. This process effectively merges contextual information from both annotations and text, thereby providing a more semantically rich input to the model.

4.5 Experiments

In this chapter, I evaluate the performance of the ECNet framework across skeleton control and facial landmark control tasks. Additionally, I demonstrate the generality of *DCL* through its application in Canny control tasks. In Section 4.5.1, the results indicate that our method surpasses the state-of-the-art methods based on SD for the multiple conditional control tasks. In Section 4.5.2, I further conducted ablation studies on *SGI* and *DCL*.

4.5.1 Comparison with SD-based Methods

4.5.1.1 Skeleton Control Task

I utilize two state-of-the-art SD-based models as the foundation for training ECNet: HumanSD (SD2.1) [105] and pose-conditioned ControlNet (SDXL) [52]. The objective function similarly employs L1 loss, here measuring the distance between facial landmarks in the generated and input images to enforce accurate facial feature generation. The training process is as follows: (1) I train HumanSD-based ECNet (ECNet (HumanSD)) using the LAION-Human dataset introduced in HumanSD; (2) I fine-tune ControlNet-based ECNet (ECNet (SDXL)) on high-resolution images (greater than 1024 pixels) selected from both the LAION-Human and HumanArt [12] datasets, ensuring the data quality aligns with our fine-tuning objectives. I benchmark ECNet models against HumanSD (SD2.1), ControlNet (SD2.1), and ControlNet (SDXL). For the sake of fairness, I retrain HumanSD on the LAION-Human dataset. The reported metrics are based on the Human-Art validation set, which contains 4,750 images.

I validate the performance of the ECNet in terms of posture generation quality and semantic association. I employ five metrics to evaluate the model: the distance-based Average Precision (AP), representing the similarity in human keypoint distances; the Pose Cosine Similarity-Based AP (CAP), indicative of human posture similarity; the People Count Error (PCE), reflecting the accuracy of generated human figures; CLIPSIM, measuring the relevance between image information and textual descriptions; and the Fréchet Inception Distance (FID), which evaluates the quality of image generation.

The results are shown in the left part of Table 4.1. ECNet demonstrates the highest similarity in human keypoint distances, reflected in its AP and CAP scores, surpassing ControlNet and HumanSD. This indicates ECNet's superior ability to capture and replicate the spatial configuration of human figures. Additionally, its lowest PCE value highlights its precision in generating multi-human figures. Our method does not supervise image quality and semantic relevance, it is likely that FID and CLIP scores may not improve. However, I achieved a significant improvement in control metrics with only a minor sacrifice in the FID and CLIP scores, demonstrating that our method is reasonable and effective. More qualitative comparisons with HumanSD and ControlNet are illustrated in Figure 4.4. The figure displays ECNet's adaptability in single and multiple-pose conditional generation.

Figure 4.4: Generated images using various SD-based models on the skeleton control task.

4.5.1.2 Facial Landmarks Control Task

Similar to the skeleton control task, I adopt ECNet (HumanSD) and ECNet (SDXL) for the facial landmark control task. The objective function is constructed using L1 loss on the landmark heatmaps from both the generated and the input images, providing supervision during the training process. I train both ECNet (HumanSD) and ECNet (SDXL) on the high-quality FFHQ [130] facial dataset. Typically, since there is no publicly available landmark-based ControlNet SDXL pre-trained model, I first trained a ControlNet (SDXL) on the FFHQ and then used it as the base model to fine-tune ECNet (SDXL). I benchmark ECNet models against HumanSD (SD2.1), ControlNet(SD2.1), and ControlNet(SDXL). To ensure fairness, HumanSD is retrained on the FFHQ dataset. The metrics are based on the WFLW [131] validation set containing 2,500 images.

I evaluated the performance of ECNet in terms of accuracy of landmarks, semantic relevance, and image quality. The model's performance is assessed using three evaluation metrics: Normalized Mean Error (NME) based on distance for assessing the accuracy of generated faces, CLIPSIM for semantic relevance, and FID for image quality. As illustrated in the right part of 4.1,

the NME scores indicate that ECNet achieves significantly higher accuracy in the facial landmarks than the baseline, suggesting our training strategy is equally effective for controlling facial generation tasks. Moreover, the reduction in FID scores in the SD2.1-base models highlights the enhanced generation quality achieved by ECNet. More qualitative comparisons are shown in Figure 4.5, comparing with ControlNet and HumanSD, ECNet showcases superior performance in facial landmarks control tasks, balancing precise control with high-quality generation.

Figure 4.5: Generated images using various SD-based models on facial landmarks control task.

4.5.1.3 Sketch Control Task

In this section, I validate the effectiveness of the ECNet in the sketch control task. I extract 90 points from strokes per sketch as annotations to serve as input for the *SGI* module. Finally, I sample 90 points from the attention distribution of the image generated by the model, aligned with the shape of the input annotations, used for *DCL* calculation in ECNet training. Qualitative comparisons with ControlNet and HumanSD, as shown in Figure 4.6, demonstrate that ECNet outperforms previous SD-based models in sketch-control generation.

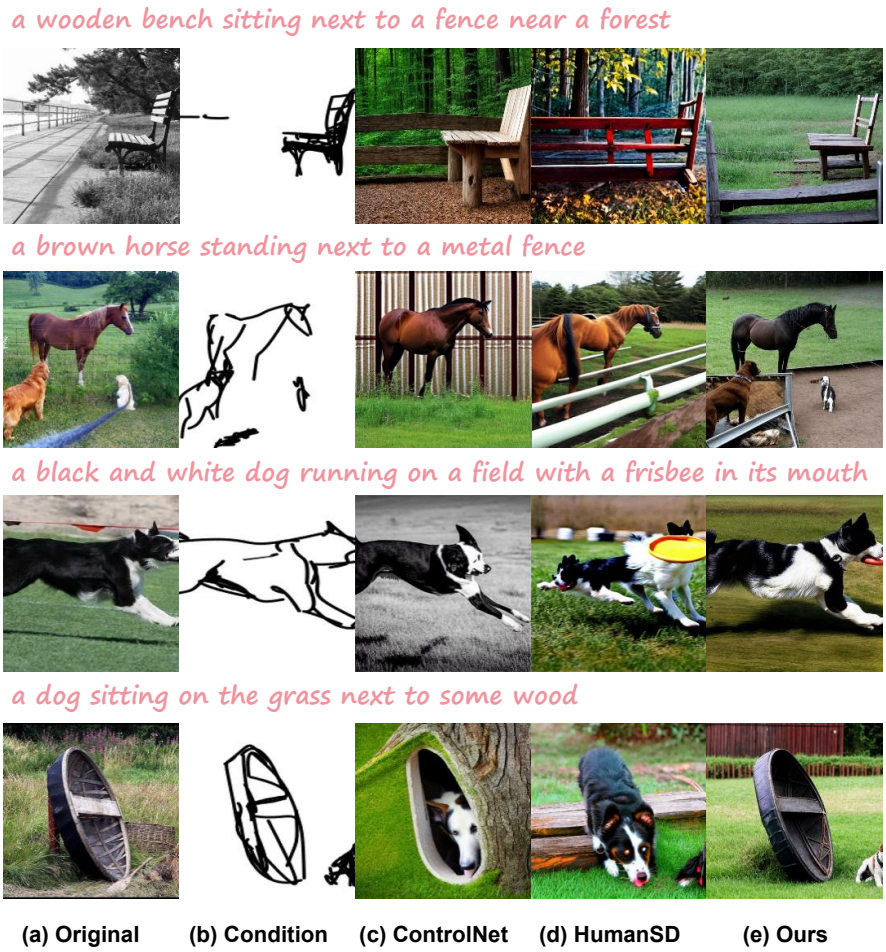


Figure 4.6: Generated images on the sketch control task. The comparison of generated results based on sketch control validates ECNet surpasses former SD-based models in this task.

4.5.1.4 Canny Control Task

In this section, I validate the versatility of the DCL loss in a control task without parametric signals, specifically using the Canny task.

I fine-tuned ECNet (SD1.5) & ECNet (SDXL) based on ControlNet (SD1.5) and ControlNet (SDXL) separately. The objective function calculates the difference between the generated and input Canny edge magnitude maps using L1 loss. I fine-tuned the ECNet (SD1.5) using a publicly available subset of the MSCOCO dataset [132], and ECNet (SDXL) using a high-resolution subset (greater than 1024 pixels) from the Laion dataset. I benchmark ECNet models against T2i-Adapter, ControlNet (SD1.5), Con-

Model	F1 Score(%)↑	CLIPSIM↑	FID↓
T2i-Adapter(SD1.5)	19.16	**32.36**	**1.75**
ControlNet(SD1.5)	25.52	32.29	2.19
ControlNet++(SD1.5)	28.15	32.16	3.72
ControlNet(SDXL)	40.24	31.93	4.64
ECNet(SD1.5)	30.96	31.72	3.21
ECNet (SDXL)	**47.13**	31.64	4.69

Table 4.2: Quantitative comparisons of Canny control tasks across different SD-based models.

trolNet++ (SD1.5), and ControlNet (SDXL). To ensure fairness, the ControlNet++ model was retrained on the same training set. For validation, I use a set of 1,800 randomly selected image-text pairs from the MultiGen20M [133] dataset.

I evaluate the performance of *DCL* in terms of edge accuracy, semantic relevance, and image quality. I use the standard metrics F1 score to assess edge accuracy, CLIPSIM to measure semantic relevance, and FID to evaluate image quality. As shown in Table 4.2, the qualitative comparison with other SD-based models demonstrates that *DCL* significantly improves the accuracy of Canny-controlled image generation, outperforming other models while maintaining strong image-text correlation and generation quality. Additional qualitative comparisons, as shown in Figure 4.7, demonstrate that ECNet exhibits superior accuracy in edge-controlled tasks compared to other SD-based models.

4.5.2 Ablation Study

In this section, I demonstrate the effectiveness of *SGI* and *DCL*, which comprise $\mathcal{L}_{drv}$ and $\mathcal{L}_{dff}$, through the skeleton control task.

4.5.2.1 Impact of Annotation Addition

To validate the effectiveness of the annotation addition module, I jointly trained the *SGI* module with our baseline model, HumanSD. As illustrated in the second row of Table 4.3, the integration of the *SGI* module enhances AP, CAP, and PCE scores, compared to the baseline model. This improvement underscores the effectiveness of incorporating such information in boosting the performance of human pose generation tasks. Furthermore, a reduction in the FID score suggests a slight improvement in the quality of images generated following the integration of the *SGI* module. Although incorporating

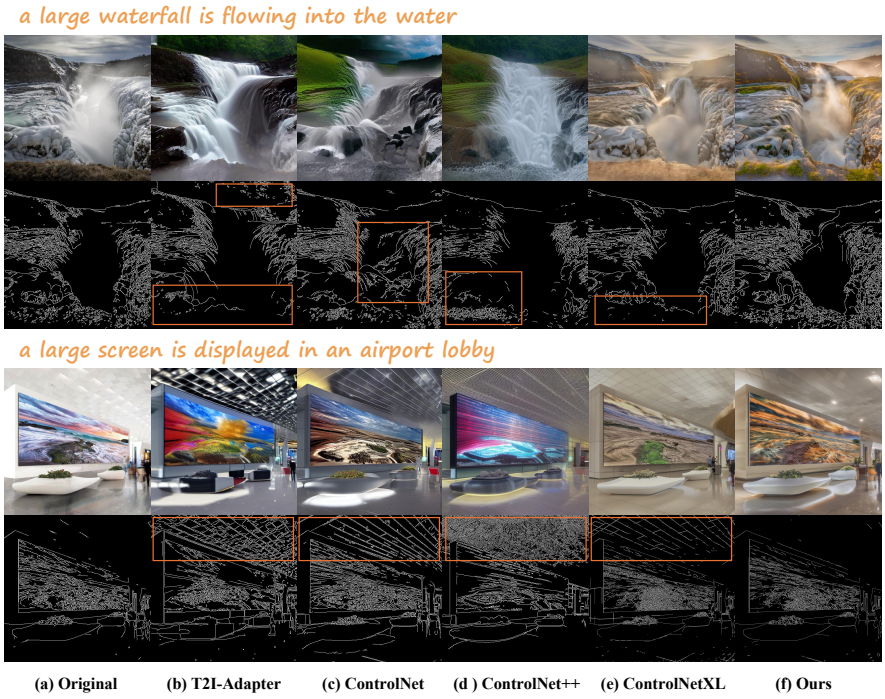


Figure 4.7: Generated images using various SD-based models on the canny control task, the areas within the orange boxes highlight regions with low control precision.

Model	AP(%)↑	CAP(%)↑	PCE↓	CLIPSIM↑	FID↓
Base	33.15	62.38	1.43	**32.63**	4.74
SGI	37.90	63.27	1.37	32.08	**4.58**
SGI&$\mathcal{L}_{drv}$	38.51	64.43	1.35	32.29	4.72
Full	**43.31**	**65.76**	**1.35**	32.28	4.89

Table 4.3: Metrics for the ablation study, performances of the base model, annotation addition, and guidance loss impact.

annotation information into text conditions does have a minor adverse effect on text features, leading to a slight decrease in the CLIPSIM index, this impact is not substantial.

4.5.2.2 Impact of different losses

In this section, I conduct validation experiments under identical conditions for the two losses proposed in the previous section: $\mathcal{L}_{drv}$ and $\mathcal{L}_{dff}$. As shown in the third row of Table 4.3, the human pose metrics of $\mathcal{L}_{drv}$

Figure 4.8: Ablation study results of base model, using the *SGI* module alone, and entire ECnet.

surpass the model with *SGI* module. It is demonstrated that supervision on denoised latent code achieves more precise control in human pose generation. Following the incorporation of $\mathcal{L}_{dff}$, as shown in the fourth row of Table 4.3, there is a further improvement in the metrics assessing human pose accuracy. This demonstrates that our proposed dual-stage loss, *DCL*, can effectively mitigate the issue of larger errors at larger time steps encountered by $\mathcal{L}_{drv}$.

Figure 4.8 illustrates the effect of applying *SGI* module and *DCL* in skeleton-based image generation. The images in the first and second rows illustrate ECNet's exceptional capability in handling scenarios involving multiple persons and in accurately generating images of rare poses. The images in the third and fourth rows highlight the efficacy of the *SGI* module in tackling the complex task of recognizing pose orientations and demonstrate how the *DCL* contributes to more precise pose control.

4.5.3 Experiment Details

All training tasks were conducted on eight NVIDIA A100 GPUs, with configurations adjusted according to the specific base models used. For HumanSD and ControlNet (versions SD1.5 and SD2.1), a batch size of 2 was employed. In contrast, ControlNet (SDXL), which has higher memory requirements, was trained with a batch size of 1. Additionally, the number of training epochs varied based on the task and dataset size. Specifically, the skeleton control task, which utilized a larger dataset, was trained for 3 epochs. Meanwhile, the facial landmarks and canny control tasks, both of which involved smaller datasets, were trained for 7 and 8 epochs respectively. This tailored approach ensured optimal utilization of computational resources and accommodated the varying complexities and data volumes of each task.

4.5.3.1 Dataset Details

In this section, I present more details of the training datasets for four tasks: skeleton control, facial landmark control, canny control, and sketch control.

Skeleton Control Task: In the experiment using HumanSD as the base model, I utilized the LAION-Human dataset curated from LAION [134], focusing on images of the highest quality that received strong approval from human evaluators. This dataset comprises 760k human image-text pairs. In the experiment using ControlNet (SDXL) as the base model, to maintain performance in generating high-resolution images, I fine-tuned the model using images from the training dataset with resolutions exceeding 1024. To augment the training data, I supplemented the LAION-Human dataset with the HumanArt dataset [12], which includes images from both natural and artificial scenes, along with clear pose and text annotations. After filtering, the combined dataset provided 4.4k image-text pairs that met the required criteria.

Facial Landmark Control Task: For this task, I used the FFHQ dataset [130] as our training set, comprising 70k high-definition facial images at a resolution of 1024*1024, representing a diverse range of ages, ethnicities, and facial attributes. For validation, I employed the WFLW dataset [131], which includes 2,500 images. I used the MMPose [2] detector to annotate each image with 98-point landmarks, and the corresponding text descriptions were generated using a pre-trained Bootstrapped Language-Image Pretraining (BLIP) model [135].

Canny Control Task: For the SD1.5-based training, I utilized a publicly available subset of the MSCOCO dataset [132], containing approximately 12k image-text pairs along with corresponding Canny edge images. For the

SDXL-based training, I selected a high-resolution subset from the LAION dataset, consisting of approximately 52k image-text pairs, and extracted Canny edge maps as the conditional images. For validation, I constructed a set of 1,800 randomly selected image-text pairs with Canny edge images from the MultiGen20M dataset [133], which comprises 20 million image-text pairs, each meticulously paired with descriptive text that encapsulates both the visual details and contextual nuances.

Sketch Control Task: In this task, I constructed a dataset using the SketchyCOCO [136] dataset. I employed CLIPasso [88], a model capable of converting images into sketches, to generate paired sketches for 5,000 image-text pairs across ten categories: airplane, bench, boat, cow, dog, elephant, horse, giraffe, train, and zebra. Dataset is a small annotated sketch-paired image-text dataset, comprising 4,000 samples for training and 1,000 for validation.

4.5.3.2 DCL details

In DCL, as shown in Equation 4.9, different loss functions are applied at various timesteps, ensuring consistent supervision across the entire diffusion process. Since different conditional control tasks have varying requirements for image fidelity, such as human keypoint detection needing only a rough pose structure, while Canny edge detection demands finer image details, the value of k is adjusted accordingly to meet these specific needs.

$$\begin{aligned} &\mathcal{L}_{DC} = \mathcal{L}_{drv} \text{ if } t < k, \text{ else } \mathcal{L}_{dff} \\ &\mathcal{L}_{drv} = |H_{inp} - H_{drv}|, \quad \mathcal{L}_{dff} = |H_{inp} - H_{dff}| \end{aligned} \tag{4.9}$$

I analyze the effects of the hyperparameter k. In different tasks, k has varying optimal values. For the facial landmark control task, the model performs best when k is set to 800. For the skeletal control task, the optimal performance is achieved when k is set to 700.
In the Canny control task, the required image fidelity for edge detection is significantly higher than that for keypoint detection. As a result, at certain ranges of timesteps, neither loss function can accurately capture the necessary edge structure, as illustrated in Fig 4.9. Through extensive validation, we found that the model achieves the best performance when $\mathcal{L}_{drv}$ is applied during timesteps 0-300, $\mathcal{L}_{dff}$ during timesteps 750-1000, and the original loss function is used in the intermediate range.

Figure 4.9: In the Canny control task, the derived image (b) and the noise difference image (c) fail to accurately extract Canny edges over a range of timesteps.

4.5.3.3 SGI Module Details

The uniqueness of SGI does not stem from its structural design but from its innovative incorporation of parametric signals derived from control conditions. Positioned behind the CLIP encoder, the SGI module integrates text features with parameter signal features through a cross-attention mechanism. In the ControlNet(SD2.1) base model, this module contains 65.2M parameters, occupying 249.8MB of memory. In the ControlNet(SDXL) base model, it expands to 201.9M parameters, with a memory footprint of 385.10MB. Despite its minimal computational overhead, the SGI module significantly enhances the model's controllability.

4.6 Conclusion

In this work, I introduce a novel framework, ECNet, built upon a pre-trained Stable Diffusion (SD) model, consisting of two main components: *Spatial Guidance Injector* and *Diffusion Consistency Loss*. I enhance the model's ability to handle ambiguities in input conditions through the introduction of *SGI*. Additionally, *DCL* is a versatile method for applying consistency supervision to the denoised latent code of the diffusion model, preserving the generative capabilities of the pre-trained SD model while amplifying the influence of various input conditions on the outputs.

One of the key challenges addressed in this work is achieving a balance between Convenience and Precision in controllable human image generation tasks. Building on the existing end-to-end SD-based framework, I developed ECNet to ensure precise alignment between model outputs and input conditions, such as poses and facial landmarks. This is accomplished through *SGI*, which mitigates condition ambiguity, and *DCL*, which enforces consistency in generation without requiring extensive iterative corrections. By maintaining high output precision while streamlining the control process,

Figure 4.10: Some failure cases with lower semantic relevance. In the facial landmark case, the generated results do not include the **red balloon** mentioned in the prompt; in the sketch case, the generated results lack the **highway** semantic.

ECNet effectively resolves this inherent tension, enabling users to achieve accurate results with minimal effort.

In our comparative analysis with base models, using diverse evaluation metrics such as posture and facial landmark accuracy, image quality, and text relevance, ECNet consistently outperforms existing state-of-the-art models across different controllable human image generation tasks. Metrics such as Average Precision (AP), Normalized Mean Error (NME), and F1 scores demonstrate significant improvements in structural accuracy and pose realism. Additionally, qualitative assessments of image quality and text relevance confirm that ECNet consistently produces contextually coherent and visually accurate character poses. These results highlight ECNet's ability to maintain high output **precision** while simplifying the user interaction process, making it an effective tool for creative tasks requiring precise yet effortless manipulability.

4.7 Limitation and Future Work

Despite our proposed ECNet enhancing controllability, it also faces certain limitations: (1) The model's supervision relies partly on detector performance, meaning annotation detection failures can impede supervisory capabilities. (2) The framework utilizes annotations as extra information added to the textual condition. This approach boosts control but lowers the relevance between image and prompt, some failure cases as illustrated in Figure 4.10. I plan to explore more robust methods of inserting annotations that balance both the model's control capabilities and semantic relevance. (3) The evaluation process remains limited, lacking thoroughness across various conditions and scenarios.

Chapter 5

Head Motion Design for Characters

Recent advances in diffusion models have greatly enhanced image and video generation quality, finding applications in virtual reality, gaming, and digital media. A key challenge in this field is generating realistic and expressive animated portraits with controlled motion, which existing methods often lack. This task highlights the contradiction between **Convenience** and **Manipulability**: achieving intuitive control overhead poses often complicates the generation process. To address this, I propose a framework that combines a trajectory-guided head pose prediction module, the Diff Transformer, with AniPortrait for generating high-quality animations driven by audio, static images, and user-defined trajectories. The Diff Transformer maps trajectory inputs to head pose sequences using Differential Attention, while AniPortrait ensures temporal consistency and visual quality. Evaluations show that our framework significantly improves pose prediction accuracy and maintains high video quality, as confirmed by user studies that highlight strong satisfaction with motion coherence and video quality. This approach enhances controllability and expressiveness, making it ideal for applications in personalized character animation, interactive media, and dynamic storytelling.

5.1 Introduction

The recent advent of diffusion models [37, 40, 42] has significantly advanced high-quality image generation. Building on this progress, several studies [72, 137–139] have integrated temporal modules, enabling diffusion models to excel in creating compelling videos. This has spurred a wave of exploration in various fields, from virtual reality and gaming to digital media, seeking new application possibilities. Among these, generating realistic and expressive portrait animations from static images has emerged as a particularly promising task. Recent research [140–142] has made notable progress in creating

smooth and natural portrait animations that maintain temporal coherence and identity consistency. However, motion control for animated characters has received limited attention within video generation. This task requires complex coordination of user inputs, identity consistency, and natural head movements to achieve visually compelling results—challenges that current methods often struggle to fully address.

To address these challenges, I propose a new framework that builds on existing diffusion models, introducing a trajectory-guided pose prediction module, the Diff Transformer, to generate high-quality animated portraits driven by multiple modalities, including audio, static character images, and user-defined trajectories. Our framework is composed of two main components. As illustrated in Figure 5.1, In the first component, I employ a Transformer-based model trained to predict head pose sequences from trajectory inputs, projecting the head movements into a 2D facial landmark sequence. The second component uses the advanced portrait video generation model, AniPortrait [140], to produce temporally consistent and realistic character animations. AniPortrait, inspired by the network architecture of AnimateAnyone [143], leverages the powerful Stable Diffusion 1.5 model to generate smooth, lifelike videos based on audio inputs, facial landmark sequences, and reference images.

Our experimental results show that this framework effectively creates character animations with natural, seamless pose transitions, high visual quality, and strong temporal consistency. By explicitly using Euler angles as intermediate features, I can seamlessly integrate custom motion sequences into the AniPortrait framework, leveraging the powerful generative capabilities of diffusion models without additional training. This approach significantly enhances our framework's versatility in facial motion editing tasks.

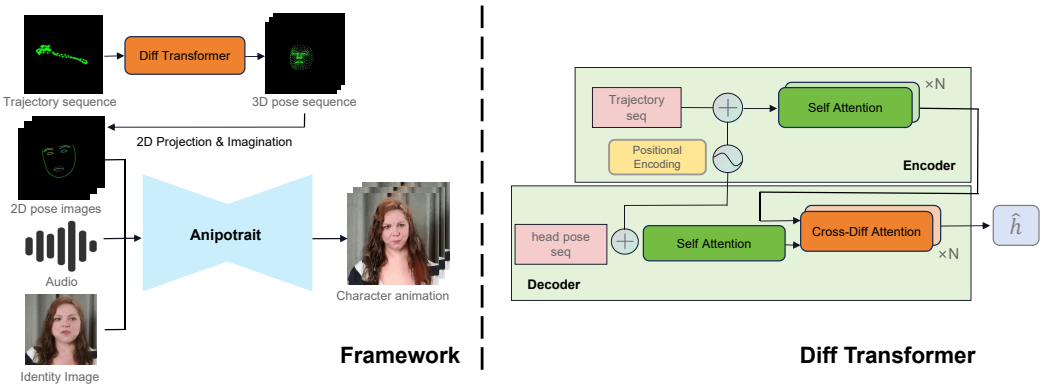


Figure 5.1: The framework of our proposed method.

5.2 Related Works

5.2.1 Video Generation with Diffusion Models

In recent years, diffusion models have achieved significant breakthroughs in text-to-image (T2I) generation [37, 40, 144, 145], which has fueled interest in extending these capabilities to text-to-video (T2V) generation [146–149], where research has increasingly focused on leveraging diffusion models for video synthesis. VideoLDM [146]introduced a motion module that employs 3D convolutions and temporal attention to capture inter-frame correlations. AnimateDiff [150] enhanced the motion modeling capability of pre-trained T2I diffusion models by fine-tuning a set of dedicated temporal attention layers on a large video dataset, allowing seamless integration with the original T2I generation process.

With advancements in powerful cross-modal encoders [27, 151], diffusion models have also been adapted in latent spaces to extend their applications to image animation tasks [37, 152]. Mahapatra et al [153] transferred estimated optical flow into artistic renderings using a pre-trained text-image diffusion model, while Li et al. [154] used diffusion models to simulate natural oscillatory motion. Several other diffusion-based approaches [72,137,139,155] have leveraged the strong generative priors of pre-trained diffusion models to achieve unprecedented open-domain animation performance. However, these models often require substantial data and time to learn complex image-to-video mappings, making the large training costs a barrier to broader accessibility.

The framework proposed in this paper offers a novel integration of custom motion cues to address the challenges of motion modeling, producing realistic and coherent video sequences without requiring model fine-tuning. This approach reduces the reliance on extensive training, making high-quality video generation more accessible to a broader range of users.

5.2.2 Controllable Video Generation with Diffusion Models

Building on the successful integration of additional conditioning signals for controlled image generation [52, 53, 133, 156], a substantial body of research [148, 157–159] has focused on incorporating diverse control signals into general video generation. These control signals include conditions on the initial video frame [148], motion trajectory [157], motion regions [158], and moving objects [159]. For instance, VideoComposer [155] introduced motion control

through the addition of motion vectors, while DragNUWA [157] generated videos conditioned on an initial image, a provided trajectory, and a text prompt. Additionally, in pursuit of high-quality video customization, some studies have explored reference-based video generation, using the motion from real videos to guide the creation of new video content [160, 161].

In current controllable video generation models, text-based controls are relatively broad and lack fine-grained control, while action-guided generation based on reference videos offers limited support for user-specific motion editing. Similarly, trajectory-based control in most approaches treats trajectories as visual features processed through CNN-based feature extraction, which primarily captures overall object motion. However, this method fails to achieve a deep coupling between the trajectory and specific motion, often leading to unnatural and disjointed movements of the subject. On the other hand, these models only focus on generating visually coherent motion sequences but lack interactive controllability, emphasizing Convenience at the expense of Manipulability.

To address these issues, I propose a Transformer-based head pose prediction module that leverages motion trajectories as spatial signal sequences, extracting attention features that dynamically relate the trajectory to the head pose sequence via a cross-attention mechanism. This approach strengthens the model's understanding of the dynamic relationship between trajectories and pose variations, enabling natural and coherent head pose adjustments driven by user-defined motion trajectories. This approach simplifies motion design while preserving interactive user control over motion trajectories, balancing Convenience and Manipulability.

5.3 Mothod

The proposed framework comprises two main modules: the Diff Transformer and a Diffusion-based video generation module. The Diff Transformer is designed to predict head pose sequences from trajectory inputs, subsequently projecting the complete head motion onto a 2D facial landmark sequence. The Diffusion-based module utilizes the state-of-the-art animated portrait generation model, AniPortrait, to produce high-quality, temporally stable character portrait videos. An overview of this framework is provided in Figure 5.1, with further details discussed in the following sections.

5.3.1 Diff Transformer

The Diff Transformer module is designed to predict head pose sequences based on input motion trajectories. This task requires the model to analyze both the motion trajectory and the preceding pose sequence, making it a natural fit for the cross-attention mechanism in Transformer models. Since the task involves understanding both broader motion trends and fine-grained details, I tailored the Diff Transformer to meet these specific demands. Inspired by prior work [3], I incorporated a specialized attention mechanism, called Differential Attention (DiffAttn), to effectively allocate attention weights across different parts of a long sequence.

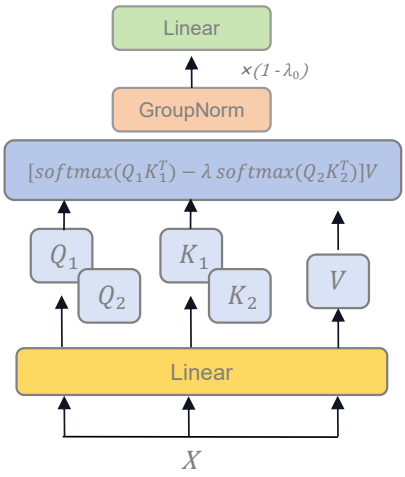
In computing cross-attention between the motion trajectory and head pose sequences, I use the DiffAttn module. This module splits the input features into two distinct query-key (QK) pairs to calculate attention weights and then computes the difference between these weights. By doing so, DiffAttn selectively enhances certain attention components while diminishing others, allowing the model to focus on local motion trends and details and thus produce natural and coherent head pose transitions. Specifically, the input X is mapped into two query sets, $Q1$ and $Q2$, and two key sets, $K1$ and $K2$, , each with d dimensions. Each query-key pair is multiplied, scaled by $s = 1/\sqrt{d}$, and processed separately to produce two attention matrices, $A1$ and $A2$. The final attention result is derived by taking the softmax-normalized difference between $A1$ and $A2$, scaled by a parameter λ. The structure of this module and its core operations are shown in Figure 5.2.

For the self-attention computation, where both sequences contribute equally in cross-attention score calculations, I retain the standard attention mechanism instead of replacing it with DiffAttn. The complete structure of the Diff Transformer is illustrated on the right side of Figure 5.1.

5.3.2 Diffusion-based Video Generation

In the diffusion-based video generation process, the model uses head poses predicted by the Diff Transformer to map facial landmarks, which serve as conditional inputs. This approach aligns head movements with audio and other input signals, enabling coherent and fluid motion throughout the animation. Reference images further ensure that each generated frame adheres to the character's identity, maintaining consistency and recognizability throughout the animation.

To enhance this approach, I integrate the AniPortrait framework, introducing a novel approach for producing high-quality animations guided by audio and a reference portrait image. The workflow involves two primary




```
def DiffAttn(X, W_q, W_k, W_v, λ):
    Q1, Q2 = split(X @ W_q)
    K1, K2 = split(X @ W_k)
    V = X @ W_v
    # Qi, Ki: [b, n, d]; V: [b, n, 2d]
    s = 1 / sqrt(d)
    A1 = Q1 @ K1.transpose(-1, -2) * s
    A2 = Q2 @ K2.transpose(-1, -2) * s
    return
      (softmax(A1) - λ softmax(A2)) @ V

def MultiHead(X, W_q, W_k, W_v, W_o, λ):
    O = GroupNorm([DiffAttn(X, W_qi, W_ki,
          W_vi, λ) for i in range(h)])
    O = O * (1 - λinit)
    return Concat(O) @ W_o
```

Figure 5.2: Structure and pseudo-code of the attention module in Diff Transformer. Right-side figure adapted from [3].

stages: first, 3D intermediate deformations are derived from the audio input and mapped into a series of 2D facial landmarks. Following this, a refined diffusion model integrated with a motion module processes these landmarks, converting them into realistic and temporally coherent portrait animations. Experimental evaluations underscore AniPortrait's advantages in delivering natural facial expressions, diverse poses, and superior visual quality, enriching the overall perceptual experience. Furthermore, AniPortrait offers flexibility and control, making it suitable for applications such as facial motion editing and face reenactment. However, while AniPortrait includes a module for inferring head movements from audio, the resulting head poses tend to be limited, lacking significant movement. This audio-driven head pose control also struggles to fully capture and represent the user's specific intent in the animation.

In our proposed framework, the audio-driven head movement is replaced by the head pose predictions from the Diff Transformer, effectively combining audio-driven lip movements with trajectory-guided head positioning. Specifically, the predicted head pose change sequence (Euler angle sequence) is merged with an expression transformation sequence derived from audio features in AniPortrait. These transformations are applied to 3D facial landmarks extracted from the reference image and then projected onto a 2D facial landmark sequence, creating an image sequence that captures the intended facial pose variations. This sequence serves as conditional input for video generation, ensuring coherent and expressive head motion in the

final animation. This integration significantly enhances the controllability and expressiveness of the generated output, aligning it closely with the intended context of the animation. Moreover, this explicit incorporation of motion information into the model is highly flexible, leveraging pre-trained diffusion-based video generation models and thus avoiding the need for extensive additional training tasks. Through the integration of various conditioning signals, the diffusion-based video generation module achieves a balance between character identity preservation and natural movement flow. This combination of flexibility and controllability makes the framework highly applicable to personalized character animation, interactive media, and dynamic storytelling applications.

5.4 Experiments

In this chapter, I evaluate the performance of the proposed framework, divided into two main sections: Transformer-based head pose prediction and AniPortrait-based video generation. The results indicate that our method outperforms the standard Transformer model baseline in head pose prediction guided by trajectory input. Additionally, our approach maintains high video quality and robustness in the video generation segment.

Table 5.1: Comparison of Pose Accuracy and FVD Scores

Pose Accuracy Comparison		
Model	**Metric**	**Error ↓**
Standard Transformer	Mean Angle Error (MAE)	1.53
Differential Transformer	Mean Angle Error (MAE)	**0.61**
FVD Score Comparison		
Method	**FVD Score ↓**	
Aniportrait	**1614.88**	
Ours	1710.17	

5.4.1 Transformer-based Head Pose Prediction

I employed an optimized Transformer model, referred to as the Differential Transformer, for head pose prediction. The objective function utilizes Mean Squared Error (MSE) loss to measure the L_1 distance between the predicted

Euler angles and the ground truth, ensuring the model accurately captures head pose changes. I adopted the Mean Angle Error (MAE) as the evaluation metric to assess the accuracy of the predicted poses. The MAE is calculated as the average difference between the predicted and ground truth Euler angle sequences, defined by the formula:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \theta_i^{\text{pred}} - \theta_i^{\text{gt}} \right|,$$

where N is the number of frames, θ_i^{pred} is the predicted Euler angle at frame i, and θ_i^{gt} is the ground truth Euler angle at frame i.

I trained the model on a single A100 GPU, Optimal performance was achieved after 100 epochs and 2 GPU hours of training. The model hyperparameters were inherited from the standard Transformer, with the maximum token length set to 150. To evaluate our model's performance, I trained both the Differential Transformer and a standard Transformer model on the same dataset comprising 1,000 videos selected from the CelebV-HQ dataset, focusing on distinct identities with clear head motion trajectories. The validation set consisted of an additional 100 video samples from the same dataset. As presented in Table 5.1, our Differential Transformer achieved an MAE of **0.61**, outperforming the standard Transformer, which achieved an MAE of 1.53. This corresponds to an improvement of approximately 60% in pose prediction accuracy, demonstrating the effectiveness of our approach.

In addition to the quantitative evaluation, I conducted a qualitative analysis by visualizing the predicted head poses overlaid on the original video frames. As shown in Figure 5.3, the Differential Transformer provides head pose estimations that are more closely aligned with the ground truth, especially in sequences with rapid head movements. This demonstrates the effectiveness of our model in capturing dynamic head pose changes more accurately than the standard Transformer.

5.4.2 AniPortrait-based Video Generation

Given the specificity of this task, there is currently a lack of directly comparable evaluation metrics. Therefore, I used Fréchet Video Distance (FVD) as an evaluation metric to assess the quality of the generated videos, and I also conducted a user study where participants rated their satisfaction with the coherence and rationality of the pose changes in the generated videos using a five-point Likert scale. I compared our model against the baseline method, AniPortrait, video inference was conducted on a single NVIDIA A100 GPU, with an average inference time of approximately 70 seconds for

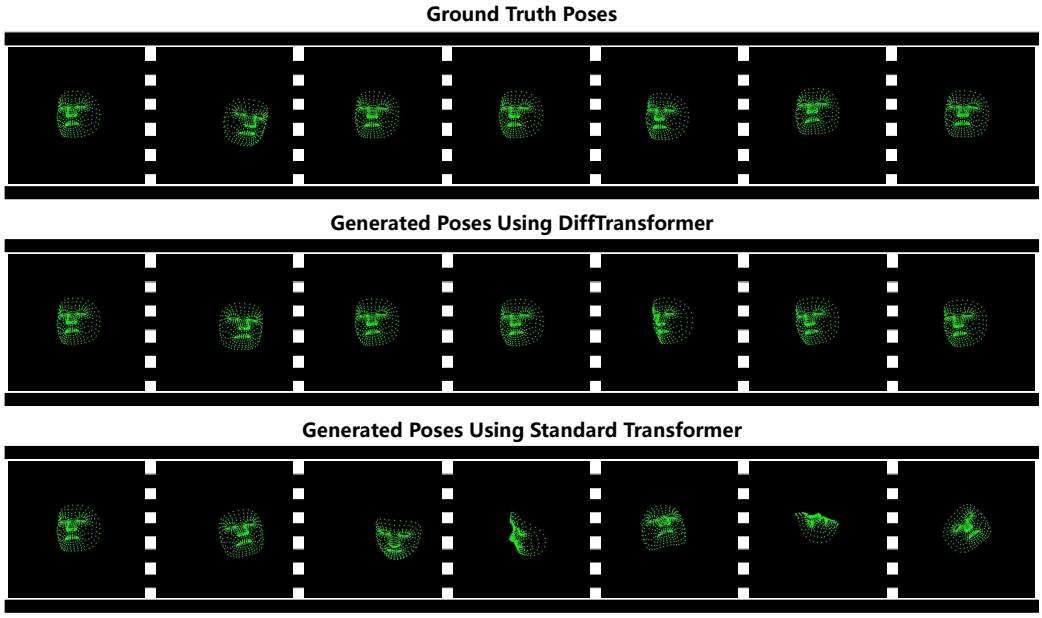


Figure 5.3: Qualitative comparison of head pose estimations from Differential Transformer and standard Transformer.

a single 4-second, 30-frame video. As indicated in Table 5.1, AniPortrait achieved a lower FVD score of **1614.88** compared to our model's score of 1710.17, suggesting that AniPortrait produces videos with slightly better overall quality as measured by FVD. In existing generative models, conflicts often emerge between geometric control and generation quality. This occurs because the introduced conditions can shift the latent space predicted by the generative model. While enhancing geometric precision may lead to a decline in FVD scores, our visual results demonstrate that the overall image quality remains high. Furthermore, the significant improvements in control achieved by our method make the minor reduction in quality an acceptable trade-off. Compared with AniPortrait, our method effectively preserves video quality and generates realistic user-controlled pose transitions, significantly enhancing the controllability of the generated animations.

To assess whether the animations generated by our proposed framework meet user expectations, I conducted a user study involving 15 participants. Each participant rated their satisfaction based on two criteria: video quality and the rationality of pose transitions. As shown in Figure 5.4, the results reveal the distribution of satisfaction ratings for both pose rationality and video quality. Most users rated the pose rationality favorably, with the highest frequency at a rating of 4, where 6 users expressed satisfaction. For video quality, the majority of users also provided positive ratings, with the

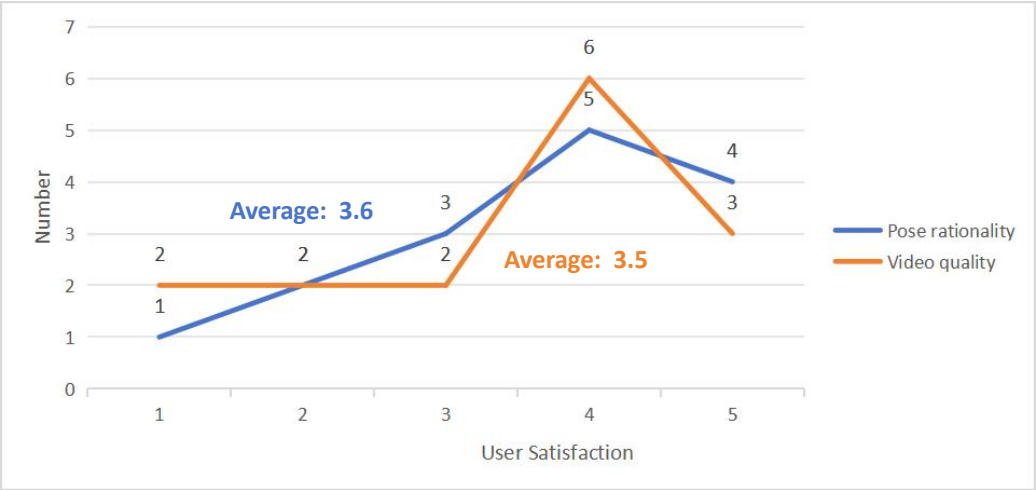


Figure 5.4: User study results on pose rationality and video quality.

peak at a rating of 3, where 5 users expressed satisfaction. Only a few participants gave lower ratings, with average scores for pose rationality and video quality at 3.6 and 3.5, respectively. This suggests that our method is generally well-received, producing coherent pose transitions and maintaining an acceptable level of video quality.

The user feedback aligns with our qualitative observations, further supporting our model's strength in generating pose transitions that appear smoother and more natural. Some examples of the generated results are shown in Figure 5.5. This balance of quantitative and qualitative evaluations highlights the effectiveness of our approach in creating videos that are not only high-quality in terms of FVD but also more engaging and satisfying from the perspective of user experience.

5.5 Conclusion

In this study, I introduced a novel framework for generating high-quality, controlled animated portraits by integrating a trajectory-guided head pose prediction module, the Diff Transformer, with AniPortrait's advanced video generation capabilities. The Diff Transformer leverages Differential Attention to translate user-defined trajectories into natural head poses, while AniPortrait maintains temporal consistency and visual fidelity. Our approach demonstrates significant improvements in pose prediction accuracy and enables seamless, expressive animations, as validated through both quantitative metrics and user studies. The feedback underscores high satisfaction with

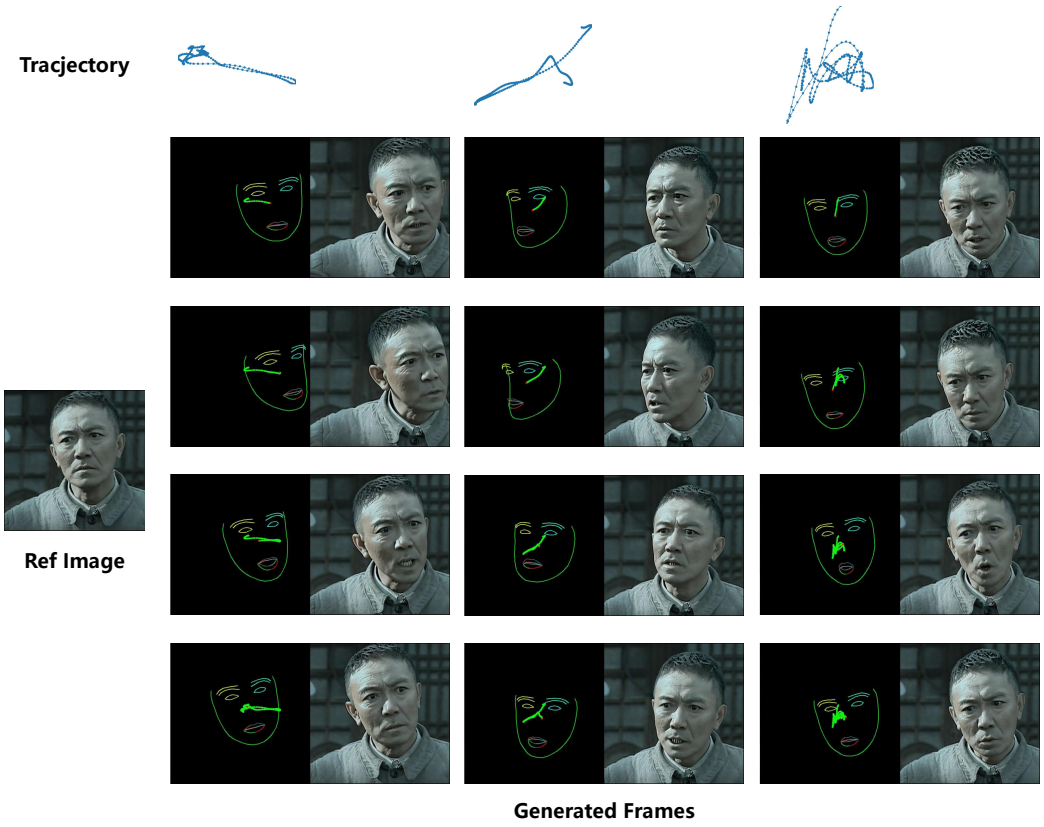


Figure 5.5: Character animations generated with varying trajectory inputs. The green curve indicates the cumulative trajectory over time, and the red point indicates the current trajectory position at the current time.

the generated motion coherence and video quality, affirming the model's capability to meet user expectations in dynamic character animation.

One of the primary challenges addressed in this work is balancing **Convenience** and **Manipulability**. While enhancing user control (Manipulability) through trajectory-guided inputs, I ensured that the system remains intuitive and efficient (Convenience) by eliminating the need for additional training or highly complex inputs. By using Euler angles and combining audio, visual, and trajectory cues, our framework achieves substantial flexibility and user control without compromising usability, effectively resolving the tension between these two elements.

The experimental results further verify the effectiveness of the system. Automated frame-by-frame head motion generation removes the need for manual animation, greatly simplifying the design process. Comparative evaluations using Mean Absolute Error (MAE) and user feedback on motion

realism confirm that generated results are both natural and reliable. Introducing trajectory point sequences as a control input extends the system's controllability within the generative framework. Metrics such as Fréchet Video Distance (FVD) and video quality scores from user evaluations demonstrate that increased control over motion trajectories does not compromise output quality or consistency.

5.6 Limatation and Future Works

Future research will explore the implicit integration of head pose predictions into the video diffusion model. Our current framework explicitly conditions on head poses, which relies on the robustness and performance of pre-trained models like AniPortrait. By embedding the head pose implicitly within the model, I aim to achieve smoother, more cohesive animations with fewer constraints tied to external model characteristics.

Additionally, our framework currently focuses on head motion control, and future extensions will investigate the application of this approach to full-body motion and other objects. Expanding beyond head movements to encompass more complex, holistic character dynamics—or even generalized object motion control—would enhance the versatility and potential applications of our framework, enabling broader use in areas such as virtual avatars and interactive media.

Chapter 6

Conclusion

In this chapter, I show additional user study results and analyze the experimental outcomes of the entire dissertation to validate the effectiveness and advancement of our research in character design across various design scenarios. By examining each scenario, I demonstrate that our proposed methods meet the core needs of users, enhance model performance, and make significant contributions to the field of multimodal character generation.

6.1 User Study on the methodology

While Chapter 1 provided a theoretical explanation of our proposed "trilemma" using Bayes theorem, I sought to further validate this framework from the user's perspective to establish a more comprehensive theoretical foundation. To this end, I conducted a user study involving direct feedback from potential end-users.

I invited eight evaluators with backgrounds that include art and design majors or experience in game character design to participate in in-depth interviews. In these sessions, I outlined the application scenarios for each design task to ensure that participants had a clear understanding of the context. I then introduced three key factors—Manipulability, Convenience, and Precision—and asked the participants to rate their level of concern for each factor based on their personal preferences and priorities in design tasks, with scores ranging from 0 to 10 representing low to high levels of concern.

The aggregated data were visualized using radar charts, as depicted in Figure 6.1. The visualization reveals that participants' preferences in each design scenario align closely with our optimization objectives. Specifically, users tended to prioritize the factors that our methods aim to enhance. This alignment between user preferences and our optimization goals provides empirical support for the validity of our theoretical framework.

The results of the user study significantly demonstrate the existence of the trilemma from the users' perspective and highlight the practical relevance of my research. By aligning my methodological developments with user

priorities, I ensure that the proposed solutions are both theoretically sound and practically impactful.

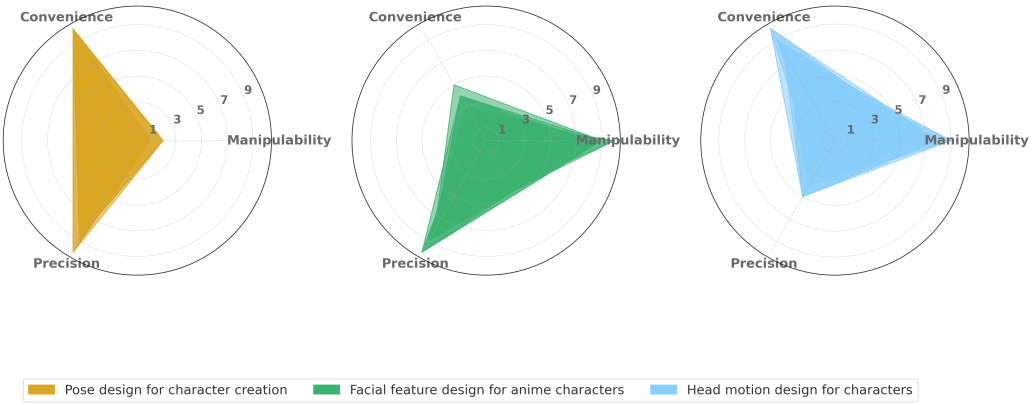


Figure 6.1: Evaluator priorities for Manipulability, Convenience, and Precision in various design tasks, visualized through radar charts. Darker regions indicate higher overlap between filled areas, highlighting the consistency and agreement in user selections.

6.2 Analysis for Performance Verification

The quantitative experiments conducted on all selected projects have strongly confirmed the improvements introduced by our methods. The significant enhancements observed in the target metrics demonstrate the effectiveness of our proposed strategies across different design tasks.

- **Drawing Multi-Age Facial Features for Anime Characters**: The highly interactive framework I developed allows users to adjust their design intentions in real-time during the creative process, significantly enhancing both **Precision** and **Manipulability**. This feature makes the framework exceptionally suitable for drawing guidance applications. Experimental results demonstrate that our model achieves an optimal balance among generation quality, diversity, and input similarity. Through autoregressive generation, it provides detailed facial features for characters of different ages, closely aligning with the user's creative intent. Consequently, our method effectively assists users in the design process of animated faces, enabling them to produce higher-quality character designs. Multi-dimensional evaluations in user studies further validate the system's effectiveness, confirming its capability to generate age-specific facial features that meet users' expectations.

- **Character Pose Design**: Based on high-precision generation results and an end-to-end framework, our approach ensures maximum **Precision** and **Convenience**—the two most critical factors for rapid character design. As evidenced by the results in Table 4.1 and 4.2, our method significantly improves the accuracy of controllable Stable Diffusion (SD)-based models across multiple control tasks. Comparative experiments and ablation studies consistently confirm the substantial enhancement in pose controllability achieved by our approach. Moreover, since our method does not rely on a specific framework, it can be flexibly integrated into various SD-based systems, further enhancing its convenience. This method offers an accurate and swift character design process, allowing users to obtain precise and vivid generated results using only simple descriptions and cross-modal input controls. Consequently, it not only elevates the quality of the generated poses but also streamlines the design workflow, enabling users to create complex character poses with minimal effort.
- **Character Head Motion Design**: Our proposed framework offers a convenient method for dynamic character animation, effectively addressing the core requirements of **Convenience** and **Manipulability** in this design scenario. By automatically generating natural and plausible motion sequences, it eliminates the need for users to engage in complex inter-frame motion editing. Simultaneously, the incorporation of trajectory input ensures that users maintain control over motion editing. As evidenced by Table 5.1, our method significantly enhances the accuracy of trajectory-based head pose predictions. Additionally, the user study results illustrated in Figure 5.4 reflect a high level of user satisfaction with the generated animations, underscoring the practicality and user acceptance of our approach in dynamic character motion design. Consequently, the framework enables the creation of natural and temporally consistent character animations using only a reference character image, thereby simplifying the design process.

6.3 Analysis for Qualitative Results

To qualitatively assess the practicality and robustness of our proposed methods, I present additional outputs from various tasks. These examples demonstrate the adaptability of our approaches to different design scenarios and highlight their potential for real-world applications. The qualitative evaluations reinforce our quantitative findings, showcasing the methods' effectiveness in producing high-quality, personalized character designs that

align with users' creative intentions.

6.3.1 Drawing Multi-Age Facial Features for Anime Characters

I conducted extensive qualitative analyses to evaluate our model's ability to generate anime characters with facial features corresponding to different ages. As shown in Figure 6.2, our model produces highly detailed and age-appropriate facial characteristics that closely adhere to the input conditions and user expectations. The generated images exhibit a natural progression of age-related features, such as changes in facial proportions and expressions, demonstrating the model's nuanced understanding of age variations in anime characters.

The comparative results highlight our model's superiority in maintaining consistency and diversity across different age groups. Unlike existing methods, our approach effectively balances the trade-off between adhering to input conditions and introducing creative variations, resulting in more authentic and engaging character designs. Users can interactively adjust inputs to fine-tune the age-specific features in real time, enhancing both precision and manipulability in the design process.

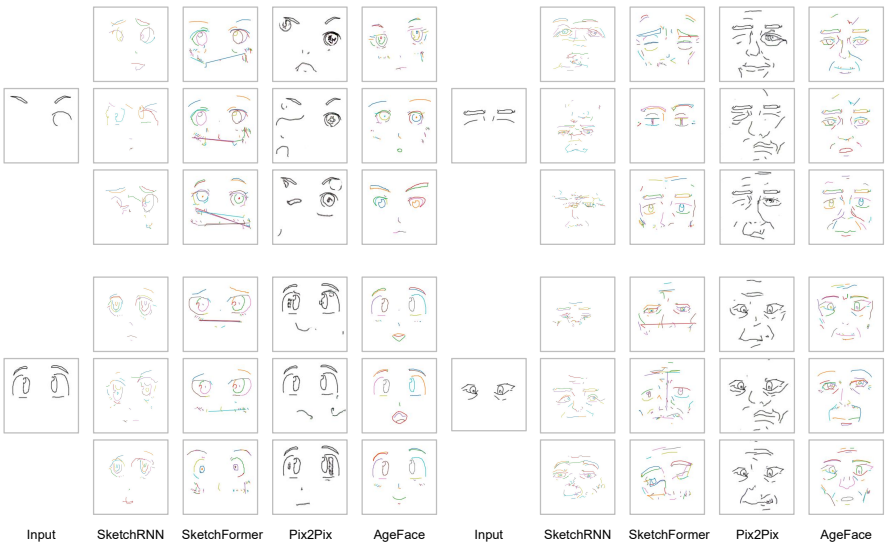


Figure 6.2: More results in comparative experiments.

6.3.2 Character Pose Design

I further assessed the qualitative performance of our method in character pose design through a series of visual comparisons. Figures 6.3, 6.4, and 6.5 illustrate our model's effectiveness in accurately replicating complex poses, facial expressions, and edge features from input conditions.

In the skeleton control task (Figure 6.3), our model consistently generates characters that precisely match the input poses, including challenging scenarios like pose orientation recognition and multi-person interactions. The generated images maintain high levels of detail and stylistic coherence, outperforming existing models in both accuracy and visual appeal.

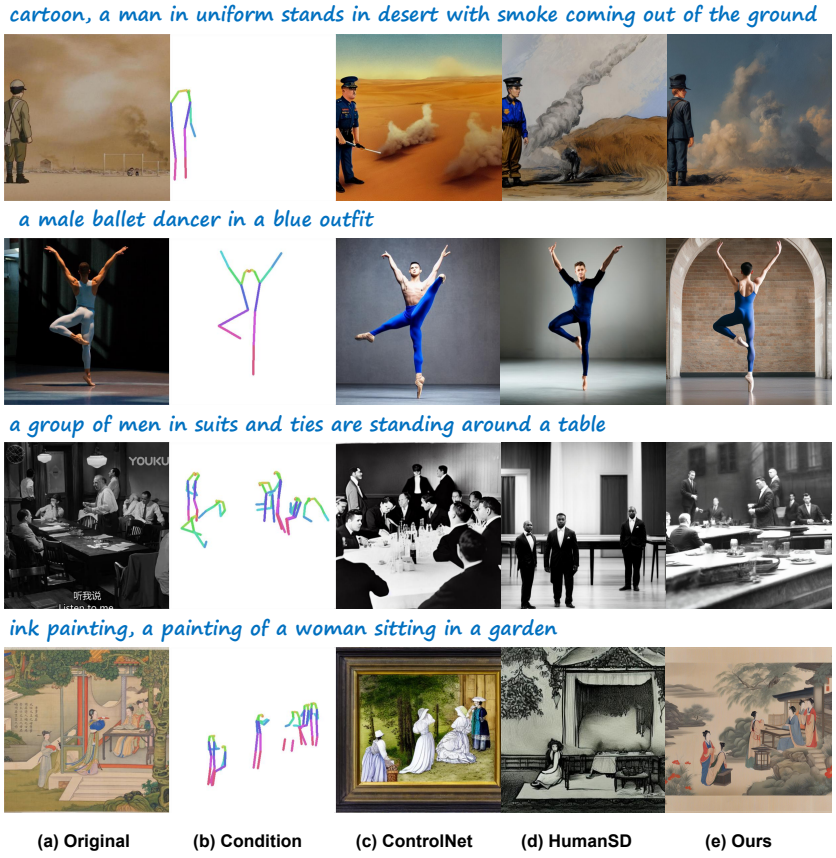


Figure 6.3: Comparison results with ControlNet and HumanSD in pose orientation recognition and multiple people scenario.

For the facial landmark control task (Figure 6.4), our method excels in capturing subtle facial expressions and nuances, resulting in highly expressive and realistic character renderings. The qualitative improvements are evident

when compared to other models, demonstrating our approach's capability to produce more lifelike and emotionally resonant characters.

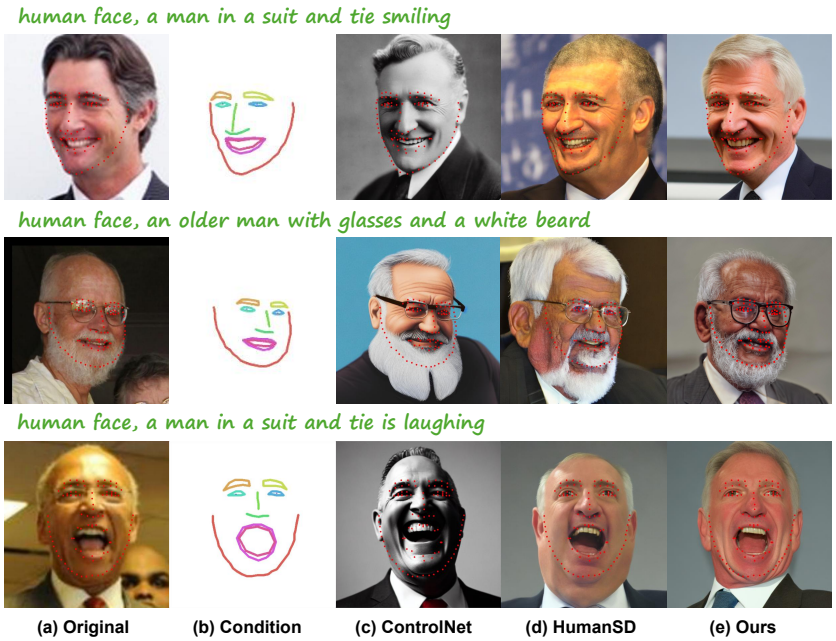


Figure 6.4: More quantitative comparison results with ControlNet and HumanSD in facial landmark control task.

In the sketch control task (Figure 6.5), our model demonstrates superior performance compared to other SD-based models. The qualitative comparisons show that our method effectively interprets and reconstructs input sketches, generating images that closely align with the user's intended design. The improved adherence to the input sketches not only elevates the quality of the generated images but also streamlines the design process by allowing for intuitive and direct control over the output through sketching.

6.3.3 Character Head Motion Design

To evaluate our method's qualitative performance in dynamic character animation, I generated additional examples in both realistic and Manga styles of character head motion designs. As depicted in Figure 6.6, our framework produces smooth and natural head movements that are temporally consistent and visually coherent.

The generated animations accurately follow the input trajectories, providing users with precise control over the motion while alleviating the need for intricate frame-by-frame editing. The qualitative assessments highlight

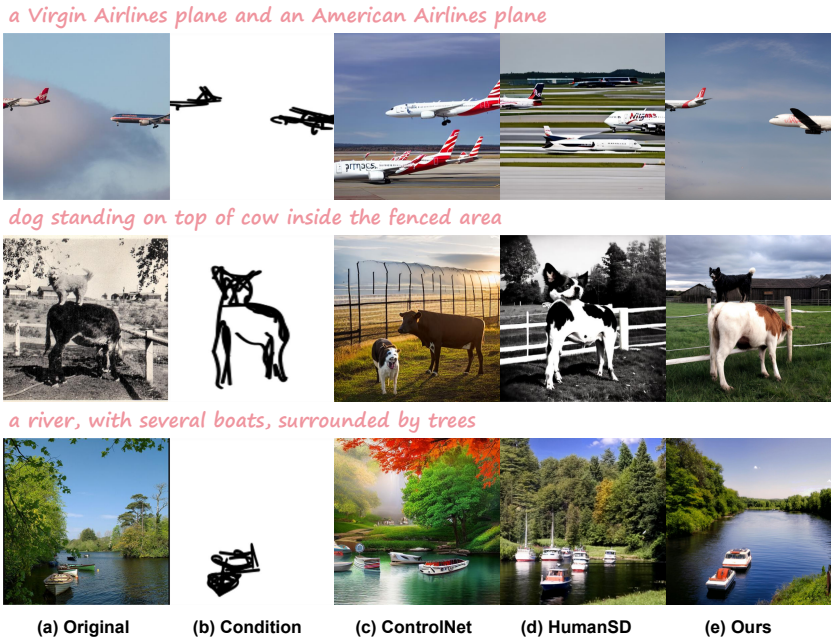


Figure 6.5: More quantitative comparison results with other SD-based models in Sketch control task.

the model's ability to maintain character identity and stylistic features throughout the animation sequence, ensuring that the motion appears both realistic and artistically pleasing.

Through the three studies presented in this dissertation, I have revealed the pervasive presence of the "trilemma" in generative models for creative design applications, highlighting the inherent tensions between Manipulability, Convenience, and Precision. By applying the methodology of primary contradictions, I demonstrated how these challenges can be systematically addressed across diverse design scenarios. Each study showcased tailored methodologies that prioritize user intent while resolving the primary conflicts unique to each application. Compared to conventional methods, our methods systematically balance the competing demands through multimodal input integration, advanced guidance strategies, and architecture-specific enhancements. This approach allows for scalable and adaptable design solutions across diverse creative tasks.

Furthermore, empirical results across three sub-tasks validate the trilemma's analysis through Bayesian marginal likelihood. Balancing accuracy and controllability in caricature generation outperformed models that were overly focused on either aspect. Similarly, achieving optimal precision in pose-controlled image synthesis required task-specific training, sacrificing gener-

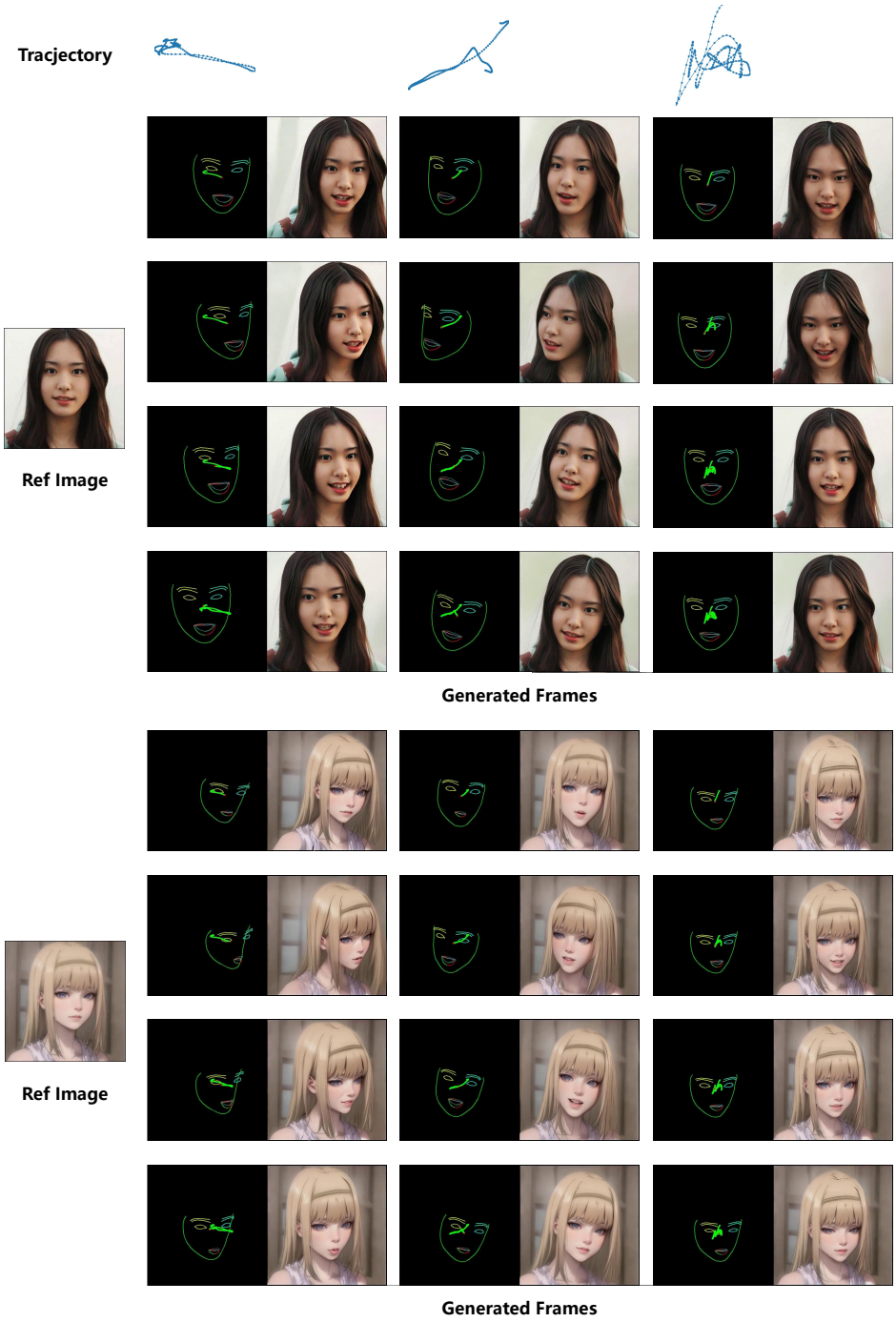


Figure 6.6: More generated character animations.

alizability. Finally, enhanced control in motion-controlled video generation negatively impacted temporal consistency due to increased generative probability distribution.

The proposed "Trilemma" thus serves as a robust methodological framework, offering valuable insights for future generative AI research in creative and design-oriented applications.

6.4 Challenges and Future Works

Despite the significant advancements achieved in this thesis, several challenges remain that present opportunities for future exploration. Enhancing the capabilities of generative models in creative design processes necessitates addressing these challenges to further bridge the gap between user intentions and generated outputs.

Firstly, while prioritizing the primary conflicting factors has been effective in resolving the "trilemma", secondary elements—though considered less critical—also play a crucial role in the overall design experience and significantly influence the user's interaction with generative models. Future research should explore methods to enhance these secondary factors without compromising the primary objectives. By finding a balance that increases the weight of secondary elements, I can develop more versatile models that offer a richer and more intuitive design experience.

Secondly, the current evaluation metrics for generative models are insufficient for accurately assessing the alignment between design intent and generated outputs. This limitation hinders the ability to quantitatively measure the performance of generative models in creative design applications, often relying on user studies with limited sample sizes, which may not yield reliable conclusions. Future work should focus on developing robust quantitative metrics that effectively evaluate the congruence between creative intent and generated results. A deeper analysis of the creative process is necessary to establish meaningful criteria and benchmarks that reflect the true effectiveness of generative models in meeting design objectives. Such metrics would facilitate more objective assessments and drive improvements in model development.

Thirdly, although this research has significantly enhanced the depth of user interaction with generative models and improved the robustness of model outputs based on creative inputs by integrating multiple modalities, the modalities considered are currently limited to digital expressions (e.g., text, audio, images). In reality, users' imagination and creative concepts are often grounded in the physical laws of the real world, which are not inherently

represented in existing generative models. This discrepancy can lead to a gap between design concepts and generated results. Future research should focus on incorporating the physical principles of the real world as a new modal within generative models. By integrating physical rules, I can more accurately reflect users' intentions and bridge the gap between virtual designs and real-world expectations, thereby enhancing the fidelity and applicability of generated outputs.

Addressing these challenges is essential for advancing the field of multi-modal character generation. By exploring ways to enhance secondary factors, developing reliable evaluation metrics, and incorporating real-world physical laws into generative models, future research can significantly improve the integration of these models into creative design workflows. This will not only expand theoretical foundations but also enhance practical applications, ultimately leading to more robust and user-aligned generative systems.

Acknowledgment

The completion of this dissertation would not have been possible without the invaluable guidance, encouragement, and support of my mentors, collaborators, family, and friends. I am deeply grateful for their contributions, which have been instrumental in helping me navigate this challenging but rewarding journey.

First and foremost, I would like to express my heartfelt gratitude to my primary supervisor, Professor Kazunori Miyata, whose unwavering support has been the cornerstone of my academic development. His insightful guidance and inspiring vision have not only shaped the trajectory of my research but also taught me the importance of perseverance and intellectual curiosity. Professor Miyata's mentorship has been a beacon of encouragement throughout my graduate career.

I am also profoundly thankful to my co-supervisor, Professor Haoran Xie, whose guidance and advice have played an integral role in my academic growth. From the very beginning of my master's studies, Professor Xie has consistently provided thoughtful feedback and practical suggestions, helping me refine my research ideas and methodologies. His expertise and continued support have been crucial to the completion of this dissertation.

I owe a special debt of gratitude to my collaborators, Keqiang Sun and Zhixin Lai, for their professional insights and selfless assistance. Their contributions and shared dedication to our projects have significantly enriched the quality of my research. Working alongside them has been a truly rewarding experience, and their collaborative spirit has greatly motivated me to achieve more.

To my family, your constant encouragement and belief in me have been my greatest source of strength. Without your support, this journey would not have been possible. I am also deeply appreciative of my friends and lab mates, who offered both intellectual and emotional support during my studies. Your camaraderie made this experience more meaningful.

To all those who have supported me in various ways, I am forever grateful. This dissertation is as much a testament to their encouragement as it is to my own efforts. Thank you!

References

[1] S.-Y. Yu, J. P. Xu, S. M. Jang, and Y. H. Pan, "A study of user experience resulting from ai artwork production - focus on character creation," *Journal of the Ergonomics Society of Korea*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268481410

[2] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[3] T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, and F. Wei, "Differential transformer," 2024. [Online]. Available: https://arxiv.org/abs/2410.05258

[4] R. Lopez, P. Boyeau, N. Yosef, M. I. Jordan, and J. Regier, "Auto-encoding variational bayes," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211146177

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139 – 144, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1033682

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219955663

[7] H.-Y. Chen, Z. Lai, H. Zhang, X. Wang, M. Eichner, K. You, M. Cao, B. Zhang, Y. Yang, and Z. Gan, "Contrastive localized language-image pre-training," 2024. [Online]. Available: https://arxiv.org/abs/2410.02746

[8] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, F. Hill, and Z. Janssen, "Multimodal few-shot learning with frozen language models," *ArXiv*, vol. abs/2106.13884, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235658331

[9] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," *2022 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5217–5227, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245123737

[10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," *ArXiv*, vol. abs/2204.14198, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248476411

[11] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *ArXiv*, vol. abs/2208.10442, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251719655

[12] X. Ju, A. Zeng, J. Wang, Q. Xu, and L. Zhang, "Human-art: A versatile human-centric dataset bridging natural and artificial scenes," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 618–629, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257365351

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1812.04948

[14] S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto, "Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing," in *International Society for Music Information Retrieval Conference*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208334750

[15] N. V. Pedro Morgado, Yi Li, "Learning representations from audio-visual spatial alignment," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *ArXiv*, vol. abs/1907.06987, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:196831809

[17] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:688013

[18] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," *ArXiv*, vol. abs/2210.08402, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252917726

[19] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *ArXiv*, vol. abs/2305.16322, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258888112

[20] C. Sun, A. Myers, C. Vondrick, K. P. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7463–7472, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:102483628

[21] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199453025

[22] H. H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201103729

[23] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *ArXiv*, vol. abs/1908.03557, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199528533

[24] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal

transformers," *ArXiv*, vol. abs/2004.00849, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214775221

[25] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8743–8752, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219617394

[26] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *ArXiv*, vol. abs/2001.07966, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210859480

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[28] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231879586

[29] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, "Florence: A new foundation model for computer vision," 2021. [Online]. Available: https://arxiv.org/abs/2111.11432

[30] H. Lin, Y. Fu, Y.-G. Jiang, and X. Xue, "Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt," 2020. [Online]. Available: https://arxiv.org/abs/2005.09159

[31] L. S. F. Ribeiro, T. Bui, J. P. Collomosse, and M. Ponti, "Sketchformer: Transformer-based representation for sketched structure," *CoRR*, vol. abs/2002.10381, 2020. [Online]. Available: https://arxiv.org/abs/2002.10381

[32] P. Xu, B. Ruan, Y. Zheng, and H. Huang, "Sketchformer++: A hierarchical transformer architecture for vector sketch representation,"

in *Computational Visual Media*, F.-L. Zhang and A. Sharf, Eds. Singapore: Springer Nature Singapore, 2024, pp. 24–41.

[33] D. P. Kingma, T. Salimans, and M. Welling, "Improved variational inference with inverse autoregressive flow," *ArXiv*, vol. abs/1606.04934, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:11514441

[34] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," 2015. [Online]. Available: https://arxiv.org/abs/1502.04623

[35] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, pp. 57:1–57:64, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208637478

[36] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *ArXiv*, vol. abs/2102.12092, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232035663

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335280

[38] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *ArXiv*, vol. abs/2105.05233, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234357997

[39] J. Ho, "Classifier-free diffusion guidance," *ArXiv*, vol. abs/2207.12598, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249145348

[40] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335086

[41] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248097655

[42] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *ArXiv*, vol. abs/2205.11487, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248986576

[43] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *ArXiv*, vol. abs/2308.06721, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260886966

[44] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 500–22 510, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251800180

[45] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ArXiv*, vol. abs/2106.09685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235458009

[46] Z. Dong, P. Wei, and L. Lin, "Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning," *ArXiv*, vol. abs/2211.11337, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:265094956

[47] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, "Pretraining is all you need for image-to-image translation," *ArXiv*, vol. abs/2205.12952, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249062786

[48] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," *ACM SIGGRAPH 2023 Conference Proceedings*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254018130

[49] A. Maungmaung, M. Shing, K. Mitsui, K. Sawada, and F. Okura, "Text-guided scene sketch-to-photo synthesis," *ArXiv*, vol. abs/2302.06883, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256846465

[50] S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Lee, and H.-Y. Tseng, "Adaptively-realistic image generation from stroke and sketch with diffusion model," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4043–4051, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251881575

[51] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "Difffacesketch: High-fidelity face image synthesis with sketch-guided latent diffusion model," *ArXiv*, vol. abs/2302.06908, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256846679

[52] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256827727

[53] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," in *International Conference on Machine Learning*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257038979

[54] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *ArXiv*, vol. abs/2302.08453, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256900833

[55] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 392–18 402, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253581213

[56] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," *ArXiv*, vol. abs/2306.10012, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259187796

[57] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, C. Xiong, and R. Xu,

"Hive: Harnessing human feedback for instructional visual editing," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9026–9036, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257622925

[58] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. teusz Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *ArXiv*, vol. abs/2005.14165, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218971783

[59] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Hu, D. Chen, and B. Guo, "Instructdiffusion: A generalist modeling interface for vision tasks," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 709–12 720, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261582721

[60] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman, "Emu edit: Precise image editing via recognition and generation tasks," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8871–8879, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265221391

[61] Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, Y. Ge, J. Zhou, C. Dong, R. Huang, R. Zhang, and Y. Shan, "Smartedit: Exploring complex instruction-based image editing with multimodal large language models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266174392

[62] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *ArXiv*, vol. abs/2304.08485, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258179774

[63] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," *2023 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pp. 18 381–18 391, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253802085

[64] Y.-Z. Song, Z. Zhang, Z. Lin, S. D. Cohen, B. L. Price, J. Zhang, S. Y. Kim, and D. G. Aliaga, "Objectstitch: Object compositing with diffusion model," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 310–18 319, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260005035

[65] K. Kim, S. K. Park, J. Lee, and J. Choo, "Reference-based image composition with sketch via structure-aware diffusion model," *ArXiv*, vol. abs/2304.09748, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258212902

[66] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," *ArXiv*, vol. abs/2210.02303, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252715883

[67] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *ArXiv*, vol. abs/2211.11018, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253735209

[68] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7312–7322, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256615582

[69] F. Liang, B. Wu, J. Wang, L. Yu, K. Li, Y. Zhao, I. Misra, J.-B. Huang, P. Zhang, P. Vajda, and D. Marculescu, "Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8207–8216, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266690780

[70] M. W. Ku, C. Wei, W. Ren, H. Yang, and W. Chen, "Anyv2v: A tuning-free framework for any video-to-video editing tasks," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268554083

[71] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *ArXiv*, vol. abs/2302.01329, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256503757

[72] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou, "I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models," *ArXiv*, vol. abs/2311.04145, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265043460

[73] C. WANG, "Ai-driven digital image art creation: methods and case analysis," *CHINESE JOURNAL OF INTELLIGENT SCIENCE AND TECHNOLOGIE*, vol. 5, no. 3, p. 406, 2023. [Online]. Available: https://www.infocomm-journal.com/znkx/EN/abstract/article_173779.shtml

[74] C. Guo, Y. Lu, Y. Dou, and F.-Y. Wang, "Can chatgpt boost artistic creation: The need of imaginative intelligence for parallel art," *IEEE/-CAA Journal of Automatica Sinica*, vol. 10, no. 4, pp. 835–838, 2023.

[75] H. JiaYi, "Analysis of chatgpt and ai painting collaborative design system under the generative artificial intelligence revolution," *Design*, vol. 08, pp. 2507–2515, 01 2023.

[76] Y. Wang, "On personalized cultural and creative product design strategy based on ai painting generation," in *2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City*, M. Atiquzzaman, N. Yen, and Z. Xu, Eds. Singapore: Springer Singapore, 2022, pp. 1317–1323.

[77] Y. J. Lee, C. L. Zitnick, and M. F. Cohen, "Shadowdraw: Real-time user guidance for freehand drawing," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011. [Online]. Available: https://doi.org/10.1145/2010324.1964922

[78] S. Li, H. Xie, and K. Miyata, "Interactive drawing interface with 3D animal model retrieval," in *International Workshop on Advanced Imaging Technology (IWAIT) 2022*, M. Nakajima, S. Muramatsu, J.-G. Kim, J.-M. Guo, and Q. Kemao, Eds., vol. 12177, International Society for Optics and Photonics. SPIE, 2022, pp. 708 – 713. [Online]. Available: https://doi.org/10.1117/12.2626086

[79] X. Du, Y. He, X. Yang, C.-M. Chang, and H. Xie, "Sketch-based 3d shape modeling from sparse point clouds," 2022. [Online]. Available: https://arxiv.org/abs/2201.11287

[80] 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.

[81] Y. Matsui, T. Shiratori, and K. Aizawa, "Drawfromdrawings: 2d drawing assistance via stroke interpolation with a sketch database," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 7, pp. 1852–1862, 2017.

[82] 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, p. 63–77, 2022. [Online]. Available: https://doi.org/10.1007/s41095-021-0227-7

[83] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg, "Quick, draw! (2019)," https://quickdraw.withgoogle.com/, 2019.

[84] N. Cao, X. Yan, Y. Shi, and C. Chen, "Ai-sketcher: A deep generative model for producing high quality sketches," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 2564–2571.

[85] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 513–520.

[86] B. Chao. (2019) Anime face dataset: a collection of high-quality anime faces. [Online]. Available: https://github.com/bchao1/Anime-Face-Dataset

[87] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[88] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir, "Clipasso: Semantically-aware object sketching," 2022.

[89] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang, "Clip2scene: Towards label-efficient 3d scene understanding by clip," 2023.

[90] S. Ge, V. Goswami, C. L. Zitnick, and D. Parikh, "Creative sketch generation," *CoRR*, vol. abs/2011.10039, 2020. [Online]. Available: https://arxiv.org/abs/2011.10039

[91] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[92] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023.

[93] T. Zhang and H. Xie, "Sketch-guided text-to-image generation with spatial control," in *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, 2024, pp. 153–159.

[94] V. team, "Sketch-a-sketch," https://vsanimator.github.io/sketchasketch/, 2023, accessed: 2024-05-27.

[95] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[96] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *CoRR*, vol. abs/1503.04069, 2015. [Online]. Available: http://arxiv.org/abs/1503.04069

[97] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[98] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, F. S. Khan, J. Laaksonen, and M. Felsberg, "Doodleformer: Creative sketch drawing with transformers," 2022.

[99] S. Ali, N. Aslam, D. Kim, A. Abbas, S. Tufail, and B. Azhar, "Context awareness based sketch-deepnet architecture for hand-drawn sketches classification and recognition in aiot," *PeerJ Computer Science*, vol. 9, p. e1186, 2023. [Online]. Available: https://doi.org/10.7717/peerj-cs.1186

[100] D. Ha and D. Eck, "A neural representation of sketch drawings," *CoRR*, vol. abs/1704.03477, 2017. [Online]. Available: http://arxiv.org/abs/1704.03477

[101] M. Shimrat, "Algorithm 112: Position of point relative to polygon," *Commun. ACM*, vol. 5, no. 8, p. 434, aug 1962. [Online]. Available: https://doi.org/10.1145/368637.368653

[102] T. Bachlechner, H. H. Majumder, Bodhisattwa Prasad Mao, G. W. Cottrell, and J. McAuley, "Rezero is all you need: Fast convergence at large depth," in *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2003.04887

[103] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, pp. 31:1–31:10, 2012.

[104] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[105] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, "Humansd: A native skeleton-guided diffusion model for human image generation," 2023.

[106] J. Piao, K. Sun, Q. Wang, K.-Y. Lin, and H. Li, "Inverting generative adversarial renderer for face reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 619–15 628.

[107] K. Sun, S. Wu, Z. Huang, N. Zhang, Q. Wang, and H. Li, "Controllable 3d face synthesis with conditional generative occupancy fields," in *Advances in Neural Information Processing Systems*, 2022.

[108] K. Sun, S. Wu, N. Zhang, Z. Huang, Q. Wang, and H. Li, "Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[109] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[110] T. Ma, B. Peng, W. Wang, and J. Dong, "MUST-GAN: Multi-level statistics transfer for self-driven person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 622–13 631.

[111] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 535–13 544.

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

[113] S. Y. Cheong, A. Mustafa, and A. Gilbert, "KPE: Keypoint pose encoding for transformer-based image generation," in *British Machine Vision Conference (BMVC)*, 2022.

[114] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," 2023.

[115] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *European Conference on Computer Vision (ECCV)*, 2024.

[116] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[117] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022.

[118] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[119] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," 2023.

[120] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," 2023.

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

[122] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7978–7986.

[123] P. Zhang, L. Yang, J. Lai, and X. Xie, "Exploring dual-task correlation for pose guided person image generation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7703–7712, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247292104

[124] L. Yang, P. Wang, C. Liu, Z. Gao, P. Ren, X. Zhang, S. Wang, S. Ma, X. Hua, and W. Gao, "Towards fine-grained human pose transfer with detail replenishing network," *IEEE Transactions on Image Processing*, vol. 30, pp. 2422–2435, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218889516

[125] S. Y. Cheong, A. Mustafa, and A. Gilbert, "Kpe: Keypoint pose encoding for transformer-based image generation," in *British Machine Vision Conference*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:260375745

[126] K. Zhang, M. Sun, J. Sun, B. Zhao, K. Zhang, Z. Sun, and T. Tan, "Humandiffusion: a coarse-to-fine alignment diffusion framework for controllable text-driven person image generation," 2022. [Online]. Available: https://arxiv.org/abs/2211.06235

[127] S. Y. Cheong, A. Mustafa, and A. Gilbert, "Upgpt: Universal diffusion model for person image generation, editing and pose transfer," *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4175–4184, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258187197

[128] J. Wang, M. Ghahremani, Y. Li, B. Ommer, and C. Wachinger, "Stable-pose: Leveraging transformers for pose-guided text-to-image generation," *ArXiv*, vol. abs/2406.02485, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270226625

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

[130] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54482423

[131] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *CVPR*, 2018.

[132] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[133] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, S. Ermon, Y. Fu, and R. Xu, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *ArXiv*, vol. abs/2305.11147, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258762776

[134] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: open dataset of clip-filtered 400 million image-text pairs," *CoRR*, vol. abs/2111.02114, 2021. [Online]. Available: https://arxiv.org/abs/2111.02114

[135] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246411402

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

[137] Y. Zhang, Z. Xing, Y. Zeng, Y. Fang, and K. Chen, "Pia: Your personalized image animator via plug-and-play modules in text-to-image models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7747–7756, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266436042

[138] Z. Dai, Z. Zhang, Y. Yao, B. Qiu, S. Zhu, L. Qin, and W. Wang, "Fine-grained open domain image animation with motion guidance," *ArXiv*, vol. abs/2311.12886, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:271325811

[139] J. Xing, M. Xia, Y. Zhang, H. Chen, X. Wang, T.-T. Wong, and Y. Shan, "Dynamicrafter: Animating open-domain images with video diffusion priors," *ArXiv*, vol. abs/2310.12190, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264306292

[140] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," 2024. [Online]. Available: https://arxiv.org/abs/2403.17694

[141] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive - generating expressive portrait videos with audio2video diffusion model under weak conditions," *ArXiv*, vol. abs/2402.17485, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268032834

[142] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, M. Huang, N. Duan, and W. Chen, "Tora: A tool-integrated reasoning agent for mathematical problem solving," *ArXiv*, vol. abs/2309.17452, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263310365

[143] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, and L. Bo, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8153–8163, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265499043

[144] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," 2022. [Online]. Available: https://arxiv.org/abs/2111.14822

[145] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Muller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *ArXiv*, vol. abs/2307.01952, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259341735

[146] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," *2023*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22 563–22 575, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258187553

[147] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. der Yang, Y. Guo, T. Wu, C. Si, Y. Jiang, C. Chen, C. C. Loy, B. Dai, D. Lin, Y. Qiao, and Z. Liu, "Lavie: High-quality video generation with cascaded latent diffusion models," *ArXiv*, vol. abs/2309.15103, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262823915

[148] H. Chen, M. Xia, Y.-Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C.-L. Weng, and Y. Shan, "Videocrafter1: Open diffusion models for high-quality video generation," *ArXiv*, vol. abs/2310.19512, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264803867

[149] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C.-L. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267028095

[150] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *ArXiv*, vol. abs/2307.04725, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259501509

[151] B. Zhang, P. Zhang, X. wen Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," *ArXiv*, vol. abs/2403.15378, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268667201

[152] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *ArXiv*, vol. abs/2010.02502, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222140788

[153] A. Mahapatra, A. Siarohin, H.-Y. Lee, S. Tulyakov, and J. Zhu, "Text-guided synthesis of eulerian cinemagraphs," *ACM Transactions on Graphics (TOG)*, vol. 42, pp. 1 – 13, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259360506

[154] Z. Li, R. Tucker, N. Snavely, and A. Holynski, "Generative image dynamics," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24 142–24 153, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261823270

[155] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *ArXiv*, vol. abs/2306.02018, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259075720

[156] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7667–7677, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261101003

[157] S.-S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *ArXiv*, vol. abs/2308.08089, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260925229

[158] Z. Dai, Z. Zhang, Y. Yao, B. Qiu, S. Zhu, L. Qin, and W. Wang, "Animateanything: Fine-grained open domain image animation with motion guidance," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265351810

[159] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, and D. Lorenz, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *ArXiv*, vol. abs/2311.15127, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265312551

[160] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J.-W. Liu, W. Wu, J. Keppo, and M. Z. Shou, "Motiondirector: Motion customization of text-to-video diffusion models," *ArXiv*, vol. abs/2310.08465, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263909602

[161] H. Jeong, G. Y. Park, and J. C. Ye, "Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9212–9221, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265608824

Publications

[1] S. Li, H. Xie, X. Yang, C. -M. Chang and K. Miyata, "A Drawing Support System for Sketching Aging Anime Faces," 2022 International Conference on Cyberworlds (CW), Kanazawa, Japan, 2022, pp. 1-7, doi: 10.1109/CW55638.2022.00010.

[2] S. Li, X. Du, H. Xie and K. Miyata, "Interactive Drawing Interface for Aging Anime Face Sketches Using Transformer-Based Generative Model," in IEEE Access, vol. 12, pp. 138751-138762, 2024, doi: 10.1109/ACCESS.2024.3466230.