Doctoral Dissertation

AMBIGUOUS-TO-CONCRETE REALISTIC HUMAN IMAGE SYNTHETIC VIA GENERATIVE MODEL

PENG Yichen

Supervisor Miyata Kazunori

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

March 2024

Abstract

In the midst of rapid advancements in data retrieval and large-scale generative model technologies, the paradigms of artistic endeavors, including writing and painting, are constantly evolving. Within the realms of Computer Graphics (CG) and Computer Vision (CV), these technological strides have substantially bolstered the efficiency of image design processes. Historically, designers would hone their artistic prowess through rigorous practice, paired with an immersion in classical artworks to cultivate their aesthetic judgment.

Generative models are able to produce complex, high-quality images from simple inputs. However, it is crucial to demystify the notion that such advancements inherently equate to significant increases in designers' productivity. While the algorithms behind AI generative models are indeed potent and efficient, their value to users is contingent upon their alignment with the designers' needs. From my perspective, the value that current image-generative models provide to designers is not commensurate with the models' performance capabilities. The majority of these models operate as end-to-end 'black boxes', making it challenging for users to achieve desired outcomes through straightforward inputs. To address this issue, this thesis discusses the following approach: 1) From the perspective of developing an algorithm: enhance the algorithm design of the generative model to allow for a variety of input modalities. This would not only improve model performance but also enrich user interaction with the algorithm. Recent advancements in multimodal generative models exemplify the capability to produce images through various inputs such as text prompts, image references, and spatial guidance, including sketches and semantic maps 2) From the perspective of interaction with models: Offer expansive interactive and editing capabilities during the creative process. By doing so, the generative model becomes a tool for exploration and refinement, rather than a one-shot solution. 3) From the perspective of the design process: It is imperative to engage with the designer's workflow, understanding the specific needs at each stage of creation. Generative model algorithms should be tailored to address these needs, ensuring that the tools developed are not just technologically advanced, but also contextually relevant.

As Picasso aptly noted, "A painting is not thought out in advance. While it is being done, it changes as one's thoughts change. And when it's finished, it goes on changing, according to the state of mind of whoever is looking at it." Thus, creation is an exploration, with designers iteratively reflecting on and drawing inspiration from intermediate outputs until satisfaction is achieved. Contemporary generative models and data retrieval techniques have yet to fully encapsulate this

ethos.

In this dissertation, we conceptualize the creative trajectory as an Ambiguous-to-Concrete continuum. Taking full-body human figure design as a case study, we break down the conventional figure painting workflow into three separate stages. For each stage, we identify the unique requirements of the user and introduce new data retrieval or generative modeling techniques specifically designed to best assist the user in achieving their creative vision. These stages include (In sequential order):

- **Posture Initialization:** At the initial stage where user intent is still forming and exploration is needed, we introduce a 'global-to-local' retrieval scheme for 3D motion data. Instead of traditional skeletal sketching, users draw trajectories for specific joints to retrieve and refine snippets of motion data, choosing keyframes that will guide further design steps. This allows users to view the motion data from different angles within a 3D space, enabling them to select the desired pose, particularly when aiming to depict dynamic actions such as dancing, thus offering users a broader range of references and choices.
- **Outfit Selection:** This phase often entails sifting through a plethora of attire references to pinpoint desired designs, coupled with iterative refinements. Considering the intricacies of fabric depiction demand profound sartorial design expertise, our solution offers an 'image-guided' generative model. This approach omits the drawing input stage, allowing users to concentrate on attire's alignment with the posture and the overall character portrayal.
- **Facial and Detail Depiction:** With the overarching attire and posture solidified, users typically possess clearer intentions regarding intricate details, especially facial attributes like hairstyle and expressions. Our solution here is a high-fidelity 'sketch-guided' generative model, ensuring the output closely mirrors the input while maintaining consistency in non-edited areas.

Through rigorous experiments, we validate the efficacy of our phase-specific interactive pipelines, benchmarking them against state-of-the-art (SOTA) counterparts on analogous tasks. The empirical results affirm that our pipeline adeptly navigates each phase of the Ambiguous-to-Concrete spectrum, offering meaningful design support. We posit that our methodologies and concepts are not confined to human figure design but are readily applicable across diverse design scenarios. As such, the frameworks and insights proffered herein can serve as foundational pillars for subsequent inquiries and innovations in the design research landscape.

Keywords: Character Image Generation, Ambiguous to Concrete, Data Retrieval, Generative Models, Artistic Creation.

概 要

データ取得と大規模生成モデル技術の急速な進展に伴い，文章表現や絵画を含む芸術活動のパラダイムは絶えず進化している．コンピュータグラフィックス（CG）とコンピュータビジョン（CV）の領域において，これらの技術的進歩はイメージデザインプロセスの効率を大幅に向上させてきた．歴史的に，デザイナーは厳しい練習を通じて芸術的能力を磨き，古典的な芸術作品に没頭することで審美眼を養ってきた．

現在，粗いスケッチやテキストプロンプトだけで，生成モデルやデータベースは魅力的な画像を生成することができ，基礎的な芸術性がほとんどまたは全くない初心者でさえも，グラフィカルなデザイン能力を発揮できるようになった．

生成モデルは，単純な入力から複雑で高品質な画像を生成することができる．しかし，このような進歩がデザイナーの生産性の顕著な向上に直接つながるかどうかを解明することは重要である．AI 生成モデルの背後にあるアルゴリズムは確かに強力で効率的であるが，その価値はデザイナーのニーズとの整合性にかかっている．筆者は，現在の画像生成モデルがデザイナーに提供する価値は，モデルの性能能力と一致していないと考える．これらのモデルの大多数はエンドツーエンドの「ブラックボックス」として機能しており，ユーザーの直接的な入力から所望の結果を得ることを困難にしている．この問題に対処するため，本論文では以下のアプローチを議論する：1）アルゴリズムの開発の観点から：入力モダリティの多様性を許容するように生成モデルのアルゴリズム設計を強化する．これにより，モデルの性能が向上するだけでなく，アルゴリズムとのインタラクションも豊かになる．最近の多様な入力，例えばテキストプロンプト，画像参照，スケッチやセマンティックマップを含む空間的ガイダンスを通じて画像を生成する能力を例証する．2）モデルとの対話の観点から：創造的プロセス中に広範なインタラクティブおよび編集機能を提供する．これにより，生成モデルはワンショットソリューションではなく，探求と洗練のためのツールとなり得る．3）デザインプロセスの観点から：創造の各段階で特定のニーズを理解し，デザイナーのワークフローに関与することが不可欠である．生成モデルのアルゴリズムはこれらのニーズに対応するように調整されるべきであり，開発したツールは技術的に優れているだけでなく，文脈的にも関連していることを確認する必要がある．

ピカソは次のような指摘をしている，「絵は事前に考えられたものではない．描かれる過程で，それは人の考えが変わるにつれて変化する．そして完成した時，それは見る人の心の状態に応じて変わり続ける．」したがって，創造は探求であり，デザイナーは満足が得られるまで中間成果に反復して思い

を巡らせ，着想を得ている．現代の生成モデルやデータ取得技術は，この理念を完全に包含しているとは言えない．

本論文では，創造的軌跡を「曖昧から具体へ」という連続体として概念化する．フルボディヒューマンイメージデザインをケーススタディとして，従来のデザインのワークフローを三つの別々の段階に分解する．各段階において，ユーザのユニークな要求を特定し，設計意図を実現するために最も適したデータ取得または生成モデリング技術を導入する．これらの段階は以下の通りである（順序に従って）：

- **ポーズ初期化：**ユーザの意図がまだ形成されておらず，探索が必要な初期段階では，「グローバルからローカルへ」という 3D モーションデータの取得スキームを導入する．伝統的な骨格スケッチの代わりに，ユーザは特定の関節の軌跡を描き，モーションデータの断片を取得する．これにより，ユーザは 3D 空間内の異なる角度からモーションデータを見ることができ，特にダンスなどのダイナミックなアクションを描写することを目指している場合に望ましいポーズを選択することができ，ユーザにより広範な参照と選択肢を提供する．
- **服装選択：**この段階では，多数の服装サンプルをふるいにかけて望ましいデザインを特定し，反復的な洗練を行うことがよくある．衣装デザインには布地に対する専門知識が必要であることを考慮して，筆者は，そのような専門知識を必要としない「イメージガイド」の生成モデルを提供する．このアプローチにより描画入力段階を省略し，ユーザが姿勢と全体的なキャラクターの描写との服装の整合に集中できるようにする．
- **顔とディテールの描写：**全体的な服装とポーズが確立されると，ユーザは通常，特に髪型や表情のような顔の属性を含む複雑な細部のデザインへと意識を移す．ここで，高忠実度の「スケッチガイド」生成モデルを提案し，出力が入力を密接に反映しつつ，編集されていない領域の一貫性を保つことを保証する．

検証実験を通じて，本研究はフェーズごとのインタラクティブパイプラインの有効性を検証し，類似のタスクにおける最先端（SOTA）の成果物と比較してベンチマークする．実証結果は，私たちのパイプラインが曖昧から具体への各フェーズを巧みにナビゲートし，意義深いデザインサポートを提供することを確認する．筆者は，提案した方法論や概念が人物像デザインに限定されるものではなく，多様なデザインシナリオに容易に適用可能であると考えている．このようにして，ここで提供されるフレームワークと洞察は，今後のデザイン研究における探求と革新のための基礎的な柱として機能することができると考える．

Keywords: キャラクター画像生成，曖昧から具体へ，データ取得，生成モデル，芸術的創造．

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The interplay between technology and artistry has always been a driving force behind innovation in the digital age. In the realm of Computer Graphics (CG), generative models for human-related content stand out as two significant domains that have transformed the creative industries, for example, fashion design and character design for video games and the metaverse. This thesis delves into the association between these two realms, exploring the potential of generative models to augment and revolutionize the human image generation process.

1.1 Background and Significance

Recently, we have witnessed unprecedented advancements in data retrieval and large-scale generative model technologies, for example, CLIP ((Contrastive Language-Image Pre-Training)) [8], StableDiffusion [6], etc. From the intelligence of search engines to content generation driven by deep learning, these progressions are reshaping how we extract knowledge from massive information and create content for daily tasks and entertainment.

Specifically in the art and design field, these technological shifts are bringing about revolutionary impacts. The research in the CG and Computer Vision (CV) field has introduced unprecedented convenience and efficiency to image design. Traditionally, designers would amount to years of practice and study to reach a high level of artistic mastery. They would delve deep into classical arts, striving to enhance their aesthetic judgments and creative techniques. Recently, a simple sketch, a descriptive text prompt, or even an abstract concept can swiftly be transformed into a concrete, high-quality image. This paves the way for creators without a traditional artistic background or professional skills to venture into new possibilities.

However, alongside these exhilarating advancements, there is a potential 'over-simplification' or 'misunderstanding' of the essence of artistic creation. Picasso's famous words remind us that art creation is not just about the end of the result; the process is even more important. Creation is not a linear, pre-determined path but a process filled with reflection, iteration, and exploration. Throughout

this process, designers continuously scrutinize their work, derive new inspirations from it, and embark on their creation once more until they are content with the outcome. The current technology, as powerful as it might be in many respects, seems to somewhat overlook the value of this exploratory, iterative process of creation.

We believe: How to integrate the irreplaceable theoretical foundation of traditional design concepts into the rapidly evolving generative algorithms? And how can we ensure that while chasing efficiency and convenience, we still retain and respect the core spirit of artistic creation? These are the significant challenges we currently face, and which serve as the foundation of our research.

In the subsequent sections of this chapter, we will delve deeper into this issue and outline our research objectives and methodology.

1.2 The Impact of CG Technology on Traditional Art and Design

The association of tradition and innovation has always been a driving force behind many of humanity's most profound achievements. Particularly in the realm of design, where the aesthetic philosophies and methodologies shaped over centuries intertwine with modern tools and techniques, it simultaneously holds immense potential and presents significant challenges.

Over time, the development of creative assistance tools has significantly impacted the artistic creation process. In the early days, traditional drawing tools like Adobe Illustrator [9] revolutionized design by enabling artists to produce vector graphics using simple geometric mathematical formulas and parameters. With just a few interactive commands, designers could effortlessly produce and modify graphical content. As technology advanced, tools like Adobe Photoshop [10] incorporated data-driven methodologies that could predict user behaviors based on their habits, effectively reducing redundant and tedious tasks. One illustrative example is Photoshop's content-aware fill, which intelligently fills in a selected portion of an image by analyzing nearby pixels. Over time, users began to acclimate to such interaction paradigms, leading to a tangible boost in design production efficiency.

Most recently, the extensive research into generative models has taken things a step further. These models not only mimic human creative actions but, in certain contexts, challenge or even surpass them. Tools like DeepArt [11] and DALL·E [1] by OpenAI [12] are prime examples of how generative models are transforming the creative landscape. They integrate machine learning and large datasets to produce artwork or design elements that might be indistinguishable

Text Prompt: *An engineer using software to create art drawing, stressed, confused.*

Text Prompt: *an robot drawing a renaissance painting, joyful, relax.*

Figure 1.1: The example for Text-to-Image generation [1].

from or even superior to human-produced content. This shift has been transformative, with more and more artists exploring and integrating these new modes of artistic creation into their workflows. Figure 1.1 shows an example of inputting a short text prompt to generate an image via DALL-E.

Modern algorithms, with their efficiency, might unintentionally eclipse the essence and value of traditional design wisdom. To this end, this paper aims to explore a new human-computer interaction creative process, which leverages the astonishing creative synthesis capabilities of today's high-performance generative models without contradicting traditional creative theories. In doing so, we hope to offer designers more possibilities in this new form of creative process.

In the following sections, we will explore the existing landscape, highlighting the strengths and shortcomings of modern generative algorithms and delve into our approach to bridge the gap between traditional design philosophies and cutting-edge technology.

1.3 Motivation

Current generative models are trained through various methods and for different objectives. Some possess robust generative capabilities, enabling them to randomly produce a wide range of high-quality samples. For instance, GANs (Generative Adversarial Networks) like BigGAN and StyleGAN2 are renowned for creating incredibly realistic images, from human faces to intricate artwork. On the other hand, some models offer great flexibility, handling inputs across multiple modalities. For example, DALL·E from OpenAI can generate images from textual descriptions, turning phrases like "a two-headed flamingo" into a

visual representation. Driven by these technological advancements, artists and designers are armed with a plethora of tools and methodologies that promise efficiency, versatility, and in some cases, astonishing creativity. However, behind this flourishing technological landscape, a subtle yet crucial aspect of the creative process seems to be overshadowed - the ambiguous-to-concrete (A2C) creative journey (we will discuss this in detail in Chapter 2.3).

Traditionally, the act of creation has always been an exploration, an intimate journey of translating the nebulous clouds of ideas and inspirations into a tangible form. This process is not linear but rather organic, allowing the creator to navigate through layers of ambiguity, continuously refining, reevaluating, and evolving the idea until it manifests into a final piece that resonates with the artist's initial vision and intent.

- **Respect for Artistic Uncertainty:** In the early stages of creation, ideas are often not fully formed. They exist in a state of fluidity and uncertainty. An A2C process allows for this natural uncertainty, enabling the artist to gradually discover and refine their vision through exploration and iteration.
- **Facilitation of Exploration:** A gradual transition from ambiguity to concreteness encourages exploration. It allows the artist to experiment with different facets of the design, fostering creativity, innovation, and the discovery of unique expressions and solutions.
- **Enhanced User Engagement:** Engaging with the design in a more phased and progressive manner can lead to a deeper connection and engagement with the creative process. It can make the process more enjoyable, meaningful, and fulfilling.

The motivation of the thesis is to intertwine these essential aspects of the traditional creative journey within the fabric of modern generative models and design tools. By doing so, our aim is to foster a more holistic, responsive, and human-centric approach to digital design and artistic creation, ensuring that technological advancements serve to enrich the creative process rather than overshadow the intrinsic human elements of artistry and creativity.

1.4 Challenges

Existing generative models, such as VAEs [13], GANs [14], and DDPMs [15], are predominantly driven by concrete objectives. Their design and training revolve around producing samples that closely emulate a target data distribution, primarily by optimizing certain loss functions. However, this approach does not always necessarily align with the creative exploration behaviors intrinsic to human artists and designers.

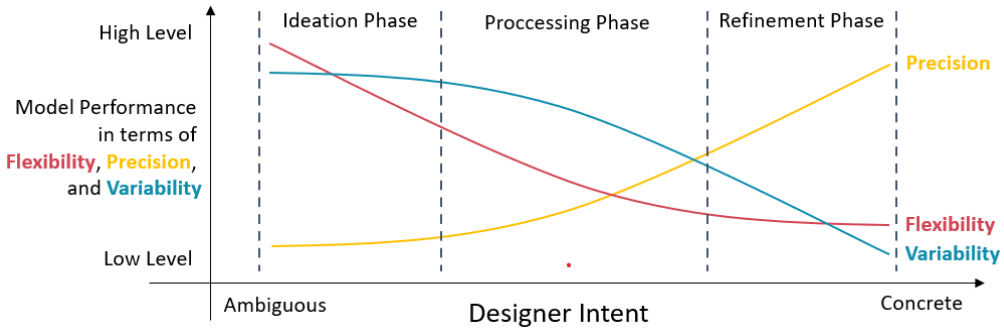

Figure 1.2: The ideal model or algorithm performance of *Precision*: Consistency between Input Conditions and Output Results, *Flexibility*: Input Compatibility, and *Variability*: Generated Samples Diversity.

Firstly, we analyze the performance of generative models across three dimensions, namely:

- **Flexibility**: This refers to the model's tolerance and understanding of vague user inputs, such as the mapping between high-dimensional abstract features across multimodalities (for example, the representation of textual information in image features, or as will be introduced in Chapter 3, the expression from abstract motion trajectories to motion and pose).
- **Variability**: This pertains to the diversity of samples that a generative model can produce. Although a higher diversity in unconditional generation tasks usually indicates a stronger learning capability of the model and a wider range of effective features learned, in conditional generation tasks, diversity often leads to generated samples deviating from the input conditions.
- **Precision**: This relates to the consistency of the samples generated by a conditional generative model with the input. Contrary to variability, higher precision in conditional generative models indicates that the model has learned a better feature mapping relationship between the input conditions and the output samples, but this can come with issues such as weak decoupling ability and low diversity of the generated samples.

Subsequently, We analyze the differing performance requirements of models at various stages of the creative process through the following examples, as illustrated in Figure 1.3. Consider a typical art classroom where a teacher prompts students to draw starting from a simple shape, say a red circle. This open-ended exercise can lead to diverse creative outcomes - one student might envision a fantastical space scene with the circle as a planet, while another could see it as

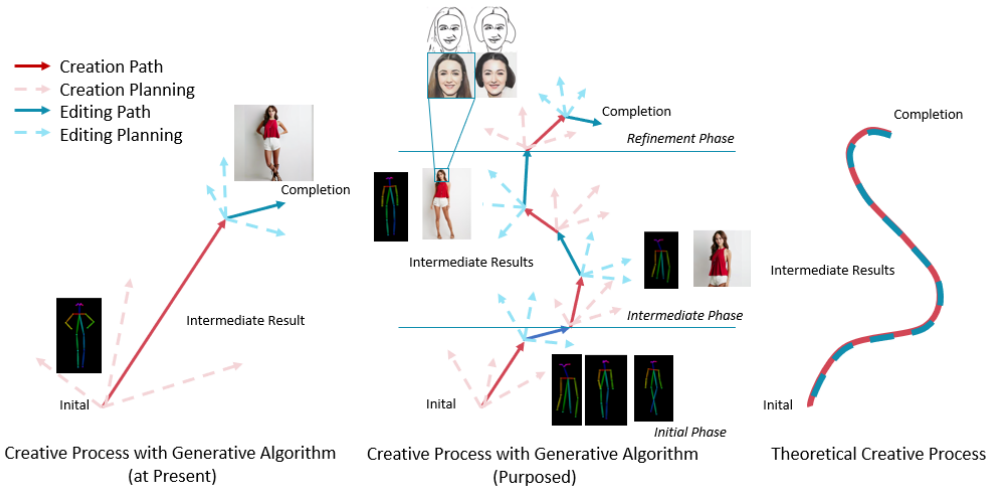

Figure 1.3: The creative process of using generative modeling (left), the proposal's creative process of using generative modeling (middle), versus the theoretical creative process (right).

the nose of a whimsical clown. The creative journey here starts from a point of ambiguity ("Draw using a circle"), refines into a more specific intent ("Use the circle as the nose of a clown"), and finally, transitions to a more detailed concept ("Sketch a clown with a specific expression and additional details").

From this analogy, it is clear that a designer's needs for a generative model vary throughout the different stages of the creative process:

- **Ideation Phase:** In the initial "Draw using a circle" phase, the artist seeks inspiration. The generative model's role here should be to understand the user's vague input and propose diverse, plausible suggestions – translating 'a red circle' into various interpretations like 'a red planet' or 'a red nose'.
- **Proccessing Phase:** Once the artist has a clearer vision, as in the "Use the circle as the nose of a clown" phase, the generative model should showcase strong generation capabilities. Given various inputs (e.g., text prompts, sketches, or image references), the model should produce outputs aligned with the theme – like images of 'a dancing clown' or 'a clown performing tricks'.
- **Refinement Phase:** Towards the end of the creative journey, when the artist is finalizing their piece, the model should offer precise editing capabilities, ensuring consistency between inputs and outputs. For instance, modifying a neutral-faced clown into one that is laughing.

1.5 Our Approach & Research Objectives

Recognizing the shifting requirements of artists and designers throughout the creative process, this thesis aims to offer a more adaptive pipeline. As shown in Figure 1.2 We classify the creative process into three distinct phases and, for each phase, formulate the ideal model or algorithm performance based on the degree of *ambiguous-concrete* of the user's intent. To address these needs, we propose three different data retrieval methods or generative models, ensuring that users can realize designs that align closely with their intentions. In this thesis, we use the design of full-body human figure images as a demonstrative example. By analyzing the traditional process of character image painting, we distilled the creation process into three distinct stages. For each stage, we take into consideration the unique needs of the users and, in response, propose three different data retrieval methods or generative models. Our goal is to more adaptively assist general users in creating designs that align with their intentions. The three stages are as follows:

- **Posture Initialization:** At this stage, user intent tends to be ambiguous and exploratory. Instead of the common approach of directly sketching the skeleton, we propose a global-to-local action data retrieval solution. Users sketch specific joint trajectories to retrieve and edit a snippet of action data from a database. This enables users to observe motion data from various perspectives in a three-dimensional space, facilitating the selection of preferred poses, especially for representing dynamic movements like dance. This provides users with an extensive array of references and options. It's akin to a photographer taking rapid succession shots of a model and then selecting the best pose from them.
- **Outfit Selection:** During this phase, users typically need to peruse a vast array of outfit references, selecting and tweaking iteratively to select the desired attire. Moreover, the rendition of fabric in art requires foundational knowledge of clothing design. Besides, during this process, the generated results require high coherence; for instance, when using a set of clothing as a reference to generate a series of images with different poses, it is necessary to maintain consistency in the clothing throughout. We introduce an image-guided generative model, eliminating the drawing input process. This allows users to focus more on the coordination of attire with poses and the overall character representation.
- **Facial and Detail Depiction:** By this stage, with the overall pose and attire of the character already determined, users have a clearer intention, especially when it comes to detailing the face (e.g., hairstyles, and expressions). We suggest a high-fidelity sketch-guided generative model. This ensures the output more closely mirrors the input, while also maintaining consistency in non-edited areas.

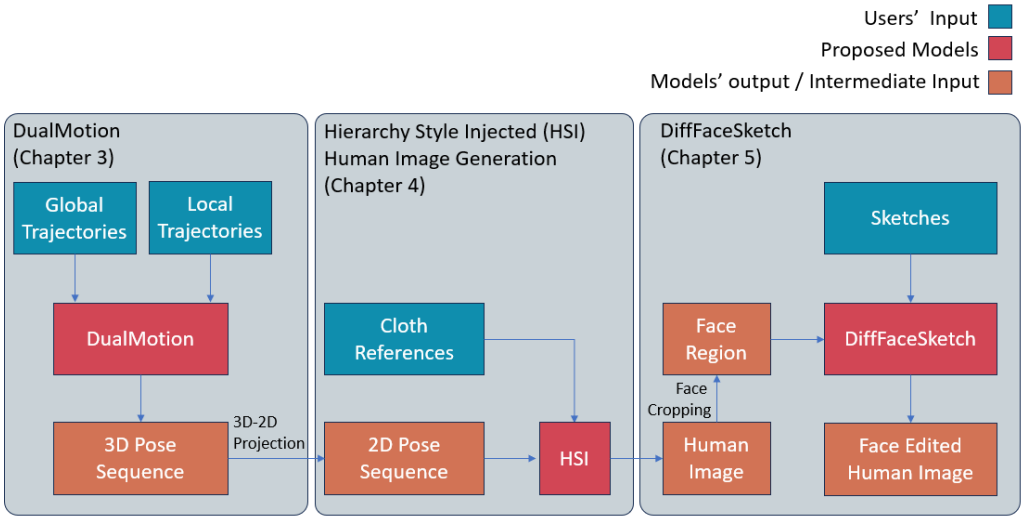

Figure 1.4: The outline of the dissertation.

Figure 1.4 illustrates the overall view of the dissertation. Our experimental validation emphasizes the efficacy of our pipeline, which prioritizes different interaction techniques at each stage. We benchmark the performance of our methods and models against state-of-the-art (SOTA) solutions for similar tasks. The results affirm that our pipeline, catering to the Ambiguous-to-Concrete creative process, worked in meaningful ways to support design. We believe that our method and concept not only apply to character image design but can also be adopted in other design creation scenarios. For instance, when creating a scene image, one should first establish the general composition, then experiment with different styles, and finally edit the local details. Or even in music composition, one would first select a few main chords, then determine the tune's style and rhythm, and finally adjust the nuances of timbre and transitions in detail. The methodologies proposed in this work serve as a theoretical foundation and offer empirical insights for subsequent studies and explorations in the field.

1.6 Organization of the Thesis

In Chapter 2, We first introduce the foundational theories on which our research assumptions are based. We then briefly review related studies in the fields of Computer Graphics (CG) and Human-Computer Interaction (HCI) where generative models have been employed to aid human creativity. We analyze the advantages and disadvantages of these approaches and articulate our research's

unique positioning within this field. Concluding the chapter, we delineate the distinctions and contributions of our research in relation to other associated studies.

In Chapter 3, we present DualMotion, a system that enables users to input motion trajectories from different viewpoints, both globally (torso) and locally (limbs). This system retrieves the most analogous skeletal motion data from a database. Not only does DualMotion streamline the editing process for character animation, but it also offers users a novel approach to specify the pose for character models.

In Chapter 4, we introduce Hierarchy Style Injection (HSI) for Human Image Generation via Diffusion Model, a methodology that leverages a pre-trained Stable Diffusion (SD) model. It is trained with a lightweight external plug-to-play module to guide the SD model's sampling. The system empowers users to control the generation of character images by inputting reference images of clothing and 2D target poses.

In Chapter 5, we delve into DiffFaceSketch, a Latent Diffusion Model specifically designed for facial generation and editing. The model ensures that edits based on user-provided sketches are faithfully represented, focusing on facial details, while maintaining consistency in non-edited regions.

Lastly, in Chapter 6, we will conclude by summarizing the principal contributions of these works, discussing their limitations, and outlining potential future research directions and planned endeavors.

Chapter 2

Related Work

Initially, we offer a brief overview of existing generative models and algorithms, analyzing their respective advantages, drawbacks, and distinctions. In Chapter 2.1, 2.2, we delve into their impact on creative endeavors. Subsequently, we present a selection of instances and studies that harness these algorithms for image generation and creative pursuits. We analyze recent advancements in the field, elucidating how contemporary work has refined and enhanced the practicality of these generative models. Finally, in Chapter 2.3, 2.4, we introduce the theoretical underpinnings of the problem setting and the associated challenges delineated from a psychological perspective. Building upon these theoretical foundations, we discuss potential enhancements to existing generative model algorithms and contemplate avenues for optimization.

2.1 Generative Model

Generative models can be broadly categorized into unconditional and conditional variants. Within the CG field, conditional generative models, in particular, have been extensively studied and employed across various content creation sectors. These models generate content by analyzing input features (conditional). Their input and output can be of the same modality, such as accepting a sketch and producing a photorealistic-styled image (tasks of this nature are termed 'image-to-image generation' or 'style transformation'). Conversely, the input and output can span different, even multiple, modalities. For instance, receiving text and audio to produce facial animations (tasks classified under 'multi-modal generation'). These various generative tasks introduce unprecedented creativity and efficiency to the realm of content creation. However, from the perspective of Chapter 1.4, there exist intriguing gaps between these generative paradigms and traditional creative processes that warrant exploration.

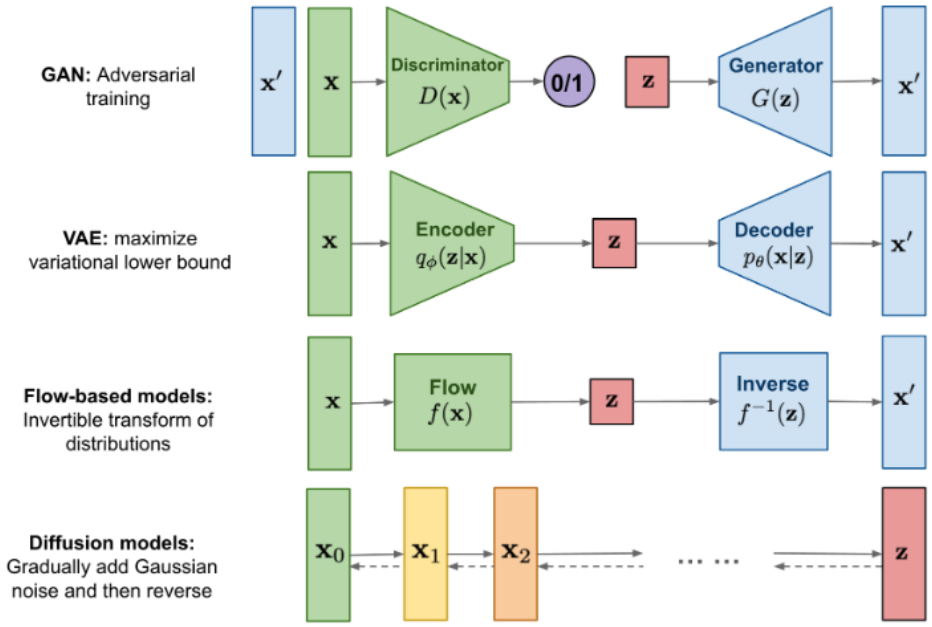

Figure 2.1: Illustrations of four distinct generative models and their sampling methods [2].

2.1.1 Architecture & Algorithm of Models

Firstly, from a structural perspective, the prevailing generative models include Variational Autoencoders (VAEs) [16], Generative Adversarial Networks (GANs) [14], flow-based models [13, 17, 18], and the recently spotlighted Diffusion Model (DDPM) [15, 19], as shown in Figure 2.1.

Mathematically, *VAEs* operate under the principle of encoding an input into a latent space and then decoding from this space to recreate the input. The objective is to optimize the likelihood of the data under a probabilistic model while regularizing the representations in the latent space using a prior (typically a Gaussian distribution). *GANs* consist of two neural networks, a generator, and a discriminator, competing in a game-theoretical framework. The generator aims to produce data that mimics real data, while the discriminator's goal is to distinguish between genuine and generated data. The process can be understood as a min-max optimization game. *Flow-based models* leverage invertible functions to transform a simple distribution (like a Gaussian) into a complex data distribution. They ensure exact likelihood computation and enable both forward and backward passes to be efficiently computed. *Diffusion models* simulate the data generation process as a reverse diffusion process, starting from a simple noise distribution and iteratively refining it until it resembles the target data distribution.

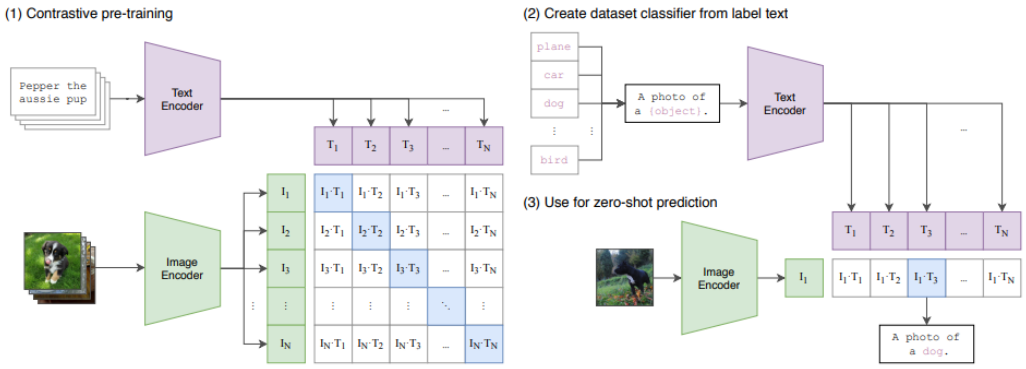

Figure 2.2: CLIP Embedding illustrations. (1) depicts how the CLIP model is trained to produce large-scale text-image embeddings through contrastive learning. (2) and (3) demonstrate the bidirectional similarity retrieval methods for text-to-image and image-to-text, respectively.

Furthermore, there exists the Contrastive Language-Image Pre-Training (CLIP) model [8]. Strictly speaking, CLIP functions as a large-scale text-to-image data retrieval algorithm. Notably, it leverages vast amounts of text-image paired data for contrastive learning to produce CLIP embeddings. As shown in Figure 2.2 (1), CLIP trains the feature similarity embedding space of text-image paired data through contrastive learning. These embeddings offer a potent prior for downstream tasks aligned with text-to-image generation. This foundational capability is evidenced in subsequently introduced models such as Stable Diffusion [6] and GLIDE [20].

Upon analysis, a commonality among these models is their pursuit of capturing the underlying data distribution, albeit through varied mathematical frameworks and objectives. While VAEs and flow-based models directly optimize data likelihood, GANs and DDPM operate under indirect measures, with GANs leveraging adversarial signals and DDPM simulating a diffusion process. While models trained on extensive data exhibit formidable sampling and generation capabilities, their generative processes can essentially be reduced to a well-defined objective-driven sampling procedure—specifically, aligning the generated sample distribution ever closer to the true data distribution. It's worth noting that some individuals, well-versed in such generative techniques, can readily discern the specific model responsible for producing a given artifact.

2.1.2 Editing Algorithm

At present, there are studies striving to ensure that the creative methods using generative models adhere, to some extent, to the phased processes that we will discuss in Section 2.3 (e.g., the loop process of creation and editing), thereby enhancing the practical utility of these generative models. Researchers have proposed various editing task algorithms for different generative frameworks. Specifically, consider a scenario where a user possesses a real sample or a generated sample from a model (such as an image of a person) and wishes to further edit it (e.g., to alter the pose of that sample). We categorize such editing algorithms into three main types: inversion, fine-tuning, and loss-guided sampling.

- **Inversion:** As exemplified by GAN Inversion [21], Textual Inversion [22] this method effectively maps a given sample back into the latent space, allowing for nuanced alterations.
- **Fine-tuning:** LORA [23, 24], Dreambooth [25], ControlNet [26], T2I-adapter [27] serves as a paradigm here, wherein the original model is subtly adjusted to produce variations of the initial output, tailored to specific user-defined criteria.
- **Attention-Guided Sampling:** The Prompt-to-promt [28], pix2pix-zero [29], Masa-control [30] model illustrates this category. It operates by steering the sampling process, leveraging loss functions to guide the generation towards desired characteristics.

One of the paramount challenges confronting these algorithms is the balancing act of maintaining content consistency and richness in detail of the input sample, while faithfully adhering to user-specified inputs (which are generally the editing based on the previous results or existing samples). For instance, in the earlier mentioned scenario, after editing the pose of the character, it becomes essential to ensure that other attributes such as physique, attire, and other distinctive features remain consistent with the original style. However, as analyzed in Section 2.1, depending on the specific phase of the creative process, users' expectations and demands from the algorithm can vary significantly. In the early stages of creation, given the often nebulous intent behind the work, users might not be overly concerned with the fidelity of the generated sample to the editing input. In fact, divergences might be viewed as a source of inspiration. Yet, as the creative process nears completion, in stark contrast, the fidelity of the generated sample to the editing input becomes paramount. An intuitive solution to this confusion, as proposed in this thesis, is to allow users to employ generative models of varying capabilities tailored to the distinct phases of their creative process.

2.2 Conversational Interfaces of Generative AI

In the preceding section, we have delineated a plethora of generative models, a compendium distilled from the confluence of traditional Computer Graphics (CG) and Computer Vision (CV) priors with advanced mathematical theories engendered by deep learning. To augment the utility of these paradigms, researchers have propounded models amenable to stylistic transmutation, denoising, super-resolution, image inpainting, and colorization. Instances such as the transformative capabilities of CycleGAN [31] and StarGAN [32] in style transfer, the noise-reducing prowess of DnCNN [33], the detail-enhancing nature of Super-Resolution CNN (SRCNN) [34], the contextually aware Generative Adversarial Networks (GANs) for inpainting, and the colorization finesse of ChromaGAN [35], are emblematic of such innovations. These models endeavor to provide users with an assortment of modal inputs and interactive schemes, thereby fostering a symbiotic nexus between the user's creative inputs and the model's generative outputs.

2.2.1 Text-to-Image Generative Model

Thanks to the powerful sequence-to-sequence architecture Transformer [36], recent developments have been particularly noteworthy with the inception of CLIP's language embeddings, which have imbued image generative models with a semblance of human-like semantic cognition, as shown in Figure 2.1. This advancement has precipitated the proliferation of various text-to-image models that have gained considerable traction within the community.

Notwithstanding, the interpretation of semantic content, a conundrum not limited to the interaction between humans and generative models but also prevalent among individuals, remains a nuanced challenge. This is not to allude to the disparities across different languages, but rather to emphasize the subjective nature of linguistic comprehension, especially with regard to adjectives, even within a singular linguistic framework. Each term, contingent on individual experiences and cognitive constructs, may lead to a spectrum of interpretations, thus rendering the task of aligning machine perception with human understanding an ongoing and formidable quest. This led to the ostensibly robust embedding space of text-to-image models to be fraught with a considerable investment in trial-and-error to fine-tune the outcome. For instance, consider the deployment of a model like DALL-E [1], which generates images from textual descriptions. The description "a two-headed eagle" can yield a multitude of visual interpretations. A user aiming for a coat of arms style representation may initially receive literal depictions of an eagle with two heads. The refinement process might entail multiple iterations,

adjusting the descriptive text to ”a heraldic two-headed eagle emblem,” to steer the model closer to the desired heraldry art style. Each iteration reflects the trial-and-error cost intrinsic to the usage of such sophisticated embedding spaces in generative models.

2.2.2 Multi-Modal Generative Model

For this reason, researchers have proposed many improvements based on the robust text-image pre-trained embeddings. Here, we briefly review some classic algorithms: ILVR [37] and SDEdit [38] were among the first to provide editable sampling solutions for image generation models based on the diffusion model, such as allowing users to mask the areas they want to edit and resample. Subsequently, methods like ControlNet [26] and T2I-adaptor [27] allowed users to input visual spatial guidance, such as line drawings, depth maps, segmentation maps, etc., in addition to text, to control model sampling. Others, such as prompt-to-prompt [28], pix2pix-zero [29], and layout-guidance [39], enable control over model sampling by allowing users to specify certain texts in text-prompts and the attention relationship with feature maps at each layer of the Diffusion Model. Additionally, the most recent works like IP-adapter [40] and MasaCtrl [30] attempt to control sample generation by directly embedding image prompts in the form of inputs into the pre-trained text-image embeddings.

These algorithms offer users more interactive methods with generative models, thereby enabling better participation and control over the generation process. In this thesis, we draw inspiration from these works and propose new algorithms for the diffusion model, tailoring the model to meet users' needs at every stage of creation and providing more reliable control and generation capabilities.

2.3 Ambiguous-to-Concrete (A2C) Creative Process

Painting serves as the foundation of all artistic creation. The designs and creations discussed henceforth take painting as a representative example. Given that the act of creation falls under the realm of art, it stands in contrast to computer science, which represents a distinct academic discipline. To my knowledge, few researchers have systematically examined human creative activity in the form of scholarly articles. However, insights gleaned from various books [41–45] and a limited number of academic papers suggest that the act of creation is an exploratory process with uncertainties. As mentioned by Hertzmann [46], the creative process is often driven by nebulous objectives, and any decision made, whether conscious or subconscious, can influence the end result. This includes choices such as the type of brush employed or the sequence in which subjects are rendered. It's worth

noting a prior reference to Picasso' s remark, 'A painting is not thought out in advance. While it is being done, it changes as one' s thoughts change...' [47] In light of these insights, I posit that a work of art never truly reaches a definitive 'completion.' The creative process endures as long as the artist identifies areas of dissatisfaction.

2.3.1 Psychological Perspectives on Creativity

Although the visual artist Chuck Close once remarked [48], 'Inspiration is for amateurs. Us professionals, we just go to work in the morning,' I find myself in partial disagreement with his sentiment. In the realm of psychology, Kahneman and Tversky [49] summarized that human behavior is predominantly governed by *intuition* (unconscious processes) and *cognition* (conscious processes). In alignment with this, Runco [50] suggested that human creative activities as being guided by *Primary processes*, characterized as 'Impulse, libido, and uncensored thoughts and feelings,' and *Secondary processes*, described as 'purposeful, rational, and guided by conventional restraints.' I categorize 'Inspiration' or 'artistic insight' as a 'passive' form of *intuition*, whereas the iterative reflection and modification during the creative process embody an 'active' form of *cognition*. These two elements operate cooperatively to guide the act of creation. Furthermore, as the creative endeavor progresses, there exists a dynamic shift in their dominance: from an initial phase where *intuition* outweighs *cognition* to later stages where *cognition* assumes a more pronounced role than *intuition*. For instance, an artist might commence with a raw, *intuitive* spark of an idea and, as the piece takes shape, rely increasingly on conscious *cognitive* strategies to refine and finalize their work. In addition, the [51] posits that, within the creative trajectory of a designer, the 'creation' and 'editing' phases should be embodied by two distinct personas. The underlying logic of this proposition suggests that the initial focus on the creative phase aligns with a more 'emotional' character, primarily driven by *intuition*. Conversely, the latter phase, emphasizing editing, necessitates a 'rational' disposition, grounded in deliberate thought and *cognition*. This further elucidates that, within the creative process, as the designer's approach to the work transitions from an 'emotional' divergence to a 'rational' rigor, there is a concomitant shift in design intent from ambiguity to specificity.

2.3.2 Creative Practice Process

As mentioned above, the creative process is an exploration of unprediction [46,52], involving an artist's skills, emotions, philosophical contemplations, and understanding of fundamental elements such as form, color, and composition. As highlighted by Adams, W. K. (2012) and Brown (2015) [51], the creative process can

be compartmentalized into four stages: *Preliminary Conceptualization*, *Planning*, *Actual Painting*, and *Reflection & Evaluation*. However, based on my personal painting experience, I discern that post the *Preliminary Conceptualization* phase, the subsequent three stages recurrently cycle until the cessation of the artwork. The intervals between these iterative loops can be as granular as each individual stroke. For instance, I might engage in *planning* whether the next stroke should delineate contours or delve into details. During the *actual painting* stage, considerations arise regarding which color or brush technique to employ. Following this stroke, an immediate *reflection* on the artwork ensues, driving decisions on whether modifications are warranted or if the creation should continue unabated. Similarly, as the creative process unfolds, the initial stages are predominantly characterized by inspiration and experimentation, typically with limited emphasis on evaluation and introspection. However, as the process matures into its latter phases, there emerges a pronounced shift, where evaluation and reflection take precedence over the earlier explorative approaches.

Drawing from the perspectives of both psychology and practical application, our analysis discerns that within the realm of artistic creation, irrespective of whether viewed from a macroscopic (entire artwork) or microscopic (individual strokes) lens, designers' creative intent is in a constant state of flux throughout their exploratory journey. In the language of deep learning models, creation (or generation) is not a linear optimization process with a well-defined objective. Instead, it can be characterized as a phase-wise fitting procedure with open-ended goals. In the sections that follow, we will enumerate pertinent algorithmic techniques and dissect the disparities and interrelationships between generative algorithms and conventional creative processes.

2.4 Experimental Human-Computer Interactive Creation Methods

While a significant majority of academic inquiries have been devoted to the enhancement and iterative improvement of algorithms, few researchers have embarked on a systematic exploration of human creative activity through the lens of scholarly publications. Even fewer studies have explored the complex interplay between AI algorithmic models and the user during the creative process. Although some of the studies cited below may not pertain directly to the Computer Graphics (CG) domain, they represent a rare subset that broaches the topic of human-AI interactivity in creative endeavors.

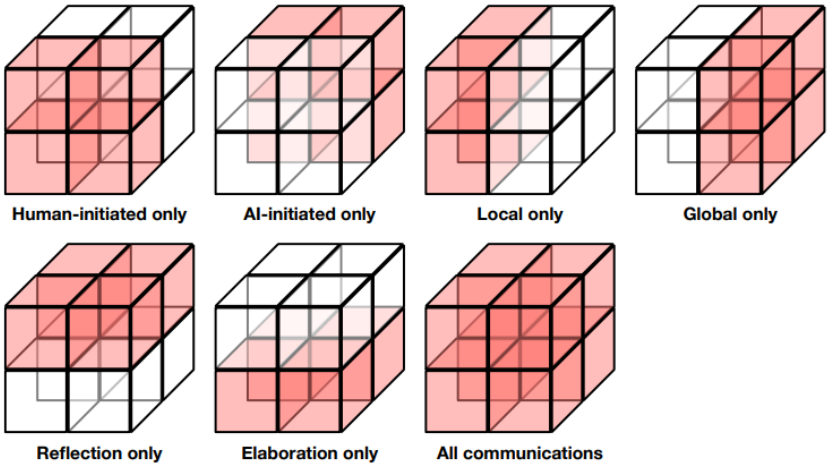

Figure 2.3: The visualized design space consists of six dimensions, which are as follows: **Human-initiated vs. Agent-initiated**, considering which party initiates the communication. **Elaboration vs. Reflection**, dealing with whether the communication pertains to previously generated content (reflection) or newly planned actions (elaboration). **Global vs. Local**, based on whether the communication addresses the creative work as a whole (global) or specific parts (local). [3]

2.4.1 Design Space

In the realm of narrative creation, [3, 53, 54] delineate a 'design space', comprising three pivotal dimensions: elaboration, reflection, and agency. 'Elaboration' underscores the system's capability to autonomously generate novel content. 'Reflection' captures the system's propensity to offer feedback and suggestions to the user. 'Agency', meanwhile, highlights the extent to which the system proactively intervenes in content modification. Furthermore, these papers introduce an additional dimension, termed 'explanation', which emphasizes the system's ability to elucidate its actions and decision-making rationales, as shown in Figure 2.3.

These concepts are also applicable in the domain of interactive generative models for computer graphics (CG), as mentioned in [55]. Given the significant divergence between generative models and human cognitive perception, to offer effective interactive schemes during the generation process, we can design a shared semantic space that both humans and generative models can comprehend, as in method [56]. This space serves as an intermediary for interaction, thereby enhancing the quality of the interaction.

Additionally, some work such as [57–59] exploring the concept of 'human-in-the-loop' (which refers to the integration of human judgment into the AI model's decision-making loop), have proposed and discussed methods that allow AI mod-

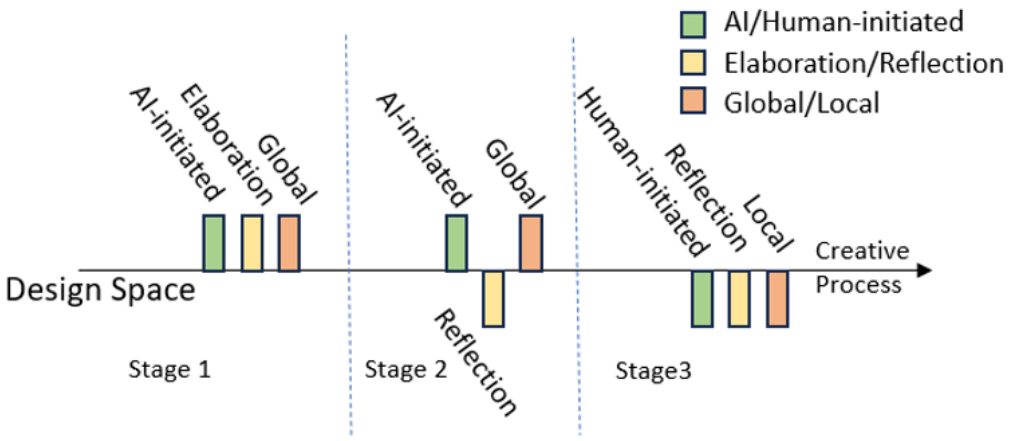

Figure 2.4: The design space is emphasized in different stages of creation.

els to learn and optimize user interaction habits through continuous interaction and iteration, thereby enhancing interaction efficiency.

With the aforementioned theoretical priors and exploratory experimental results, in this thesis, we take person image generation as an example, dividing the entire process into three stages, and categorizing the interaction preferences of these three stages within the design space, as shown in Figure 2.4. By proposing three different models, we attempt to align the generative system more closely with the Ambiguous-to-Concrete (A2C) Creative Process, thereby enhancing the rationality of interaction and the quality of generation.

Chapter 3

Global-to-Local Casual Motion Design

In this chapter, we shall delve into the intricacies of the 'Global-to-Local Causal Motion Design'—a sophisticated system that facilitates users in driving skeletal motion data outputs by inputting action trajectories from varied perspectives, both global (such as the torso) and local (like the limbs), as shown in Figure 3.1 (a). The proposed system, aside from streamlining the modification process of character animations, introduces an innovative approach for users to dictate specific postures from edited motion sequences, as shown in Figure 3.1 (b).

3.1 Introduction

Creativity support systems have gained considerable attention in computer graphics and human-computer interaction domains, catering to both professional and mainstream users. Specifically, character animation production is a prevalent creative endeavor spanning multiple sectors, including entertainment, video games, films, sports, and medical domains.

Crafting realistic character animations demands a certain level of expertise and labor-intensive processes. This complexity often deters amateur users from even attempting to generate basic character animations. In the professional realm, motion capture (often abbreviated as Mocap) remains the go-to technique to produce these authentic animations. This procedure involves one or more actors adorned with sensors performing the required motions. Here, "authentic" or "natural" implies that the animation convincingly portrays a virtual human's presence.

As these actors perform, cameras strategically placed within the environment capture and record the marker's positional data at high frequencies. However, the associated costs — encompassing labor, location, duration, and equipment — render motion capture both expensive and time-intensive. Despite the accumulation of substantial mocap data and the availability of many mocap databases (e.g., [60] and [61]), leveraging existing datasets remains challenging for novices. Tailoring animations to specific applications or achieving nuanced motion details

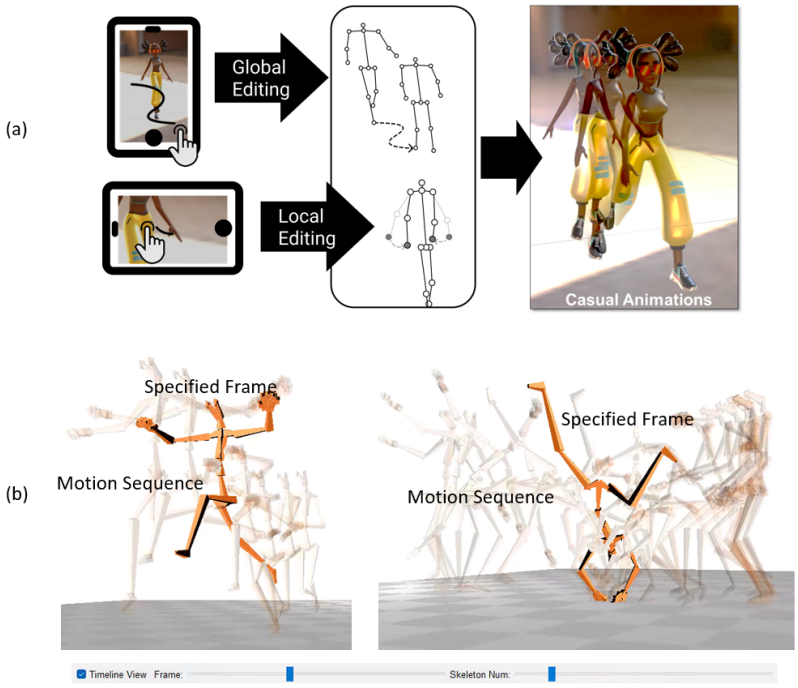

Figure 3.1: *DualMotion* enables users to input rough sketches and create casual character animations (a), dictate postures from motion sequences (b).

often necessitates recording mocap data afresh.

Additionally, beyond the design of skeleton animations, there are challenges in the design (specification) of single-frame poses as well. Some existing methods, as cited, use user-inputted 2D line drawings (simple human tracings or stick figures [62, 63] or fit static poses with real RGB photos [64, 65], as in other citations. The former requires some drawing fundamentals, like human proportions and pose balance, while the latter is not conducive to editing. Both approaches face the issue of 'requiring users to have a concrete design intent.' As mentioned in Section 2.3, when designing a human image, designers often do not have a clear initial intent for structures like poses (for example, the ambiguous idea of 'a person dancing'), and it is difficult to immediately conceive a complete pose image in one's mind (for instance, a specific dancing position: standing on one leg, hands raised high, etc.).

While many Mocap databases offer keyword-based search functionality, these subjective keywords often fail to encapsulate every nuance within a motion sequence, hindering animators from pinpointing specific motion segments within vast databases.

Alternate methodologies for motion retrieval, such as sketch-based interfaces, have been examined in prior research, like [66]. These interfaces, prevalent in computer graphics tasks like interactive animation design [67–69], 3D modeling [70], and mesh modification [71], are lauded for their intuitiveness, especially among novices.

Nevertheless, integrating sketch-based interfaces for motion retrieval presents challenges. Motion capture data, inherently complex due to its spatio-temporal characteristics, comprises a static segment (detailing the skeletal structure and node offsets) and a dynamic segment (showcasing pose variations across frames through relative node rotations). Typically, a human animation encompasses numerous nodes and potentially thousands of frames.

Conversely, the 2D sketching realm is far more straightforward dimensionally, limiting its expressiveness for intricate tasks like motion extraction and assembly. The proliferation of touchscreens and micro-video platforms further complicates character animation creation, given the limited precision and scope of such devices. This setting, where users tend to sketch roughly, magnifies the challenges associated with transforming sketches into animations.

Addressing these challenges, we introduce *DualMotion*, a tiered sketching paradigm for novice-friendly character animation retrieval and synthesis in casual animation contexts. This framework simplifies the intricate design process by segmenting it. At each phase, users focus on a specific design facet. For sketch-driven motion extraction and fusion, this process bifurcates into the global and local stages. The former pertains to broad character movements, while the latter dives deeper, focusing on individual limb movements. A data-backed methodology aids users in locating and amalgamating suitable motions. Additionally, both stages provide motion suggestions in the form of dashed polygons for user guidance.

Implementing this approach, we executed a user-centric study contrasting our dual-phase interface with a singular-phase counterpart. Empirical evidence indicates our method's superior ability (with a significance level of $p < 0.01$) in assisting beginners with motion exploration and customization.

Our key contributions encapsulate:

1. A two-tiered creative support system designed for character animation retrieval and synthesis, enabling beginners to tackle one design aspect at a time.
2. The actualization of *DualMotion*, a manifestation of our proposed framework, empowers novices to craft conventional motions using a data-driven approach.
3. A comprehensive user evaluation involving 14 participants to ascertain the efficacy and applicability of both our proposed framework and the *DualMo-*

tion tool.

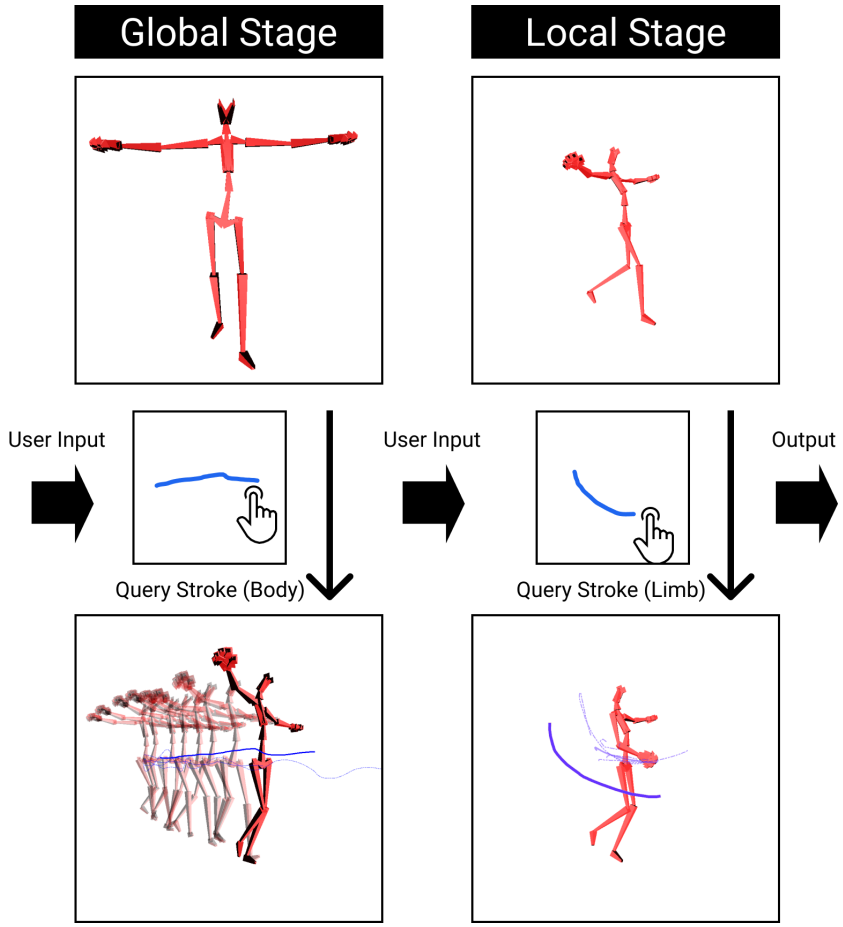

Figure 3.2: The proposed Global-to-Local design scheme for data-driven character animation design. In the global stage, users retrieve the rough motion of the whole character by inputting a query stroke; in the local stage, users further draw the strokes to edit the upper limb motion.

3.2 Related Work

This section reviews prior work on frameworks for (1) character animation authoring and (2) motion retrieval and composition.

3.2.1 Authoring Character Animations

Recent advancements have allowed for the creation of character animations through various methods such as physics-based simulations [69, 72, 73] and demonstration by users [74–76]. Despite the effectiveness of these techniques for elementary object movements, like those seen in graffiti, they fall short when applied to the complex domain of skeletal character motions. This shortfall stems from the intricate structure of skeletal data, which comprises over N joints (where $N > 20$), necessitating the user to manually adjust each joint angle frame-by-frame—a process that can be both labor-intensive and monotonous.

In the realm of animation, motion capture (Mocap) data is frequently employed as a solution to circumvent the intricacies of manual joint manipulation. Mocap data are essentially recordings of real objects or human movements, providing a foundation for realistic motion in animations. The work by Dontcheva et al. [77] exemplifies the use of a specialized Mocap system that enables animators to craft character animations by emulating captured movements. Further contributions to this field include the assembly of extensive Mocap databases by numerous research entities and commercial enterprises [60, 61], along with the proposition of various methodologies for generating motion sequences leveraging these databases [78, 79].

Our approach is data-driven, yet we place a specific focus on two pivotal aspects: (i) the intuitive retrieval of motion data from the extant databases, and (ii) the subsequent tailoring of this motion data to meet specific animation requirements. The crux of our work lies in simplifying the animation process while ensuring that the resultant motions are both natural in appearance and aligned with the animator's vision.

3.2.2 Retrieving and Editing Character Motion Data

In the quest to harness the wealth of motion data contained within expansive Mocap databases, the field has seen extensive research into retrieval techniques. For instance, keyword-based searches have been commonplace. Kruger et al. [80] paved the way with an algorithm for similarity searches, facilitating motion retrieval from large-scale Mocap databases. Complementing textual approaches, Peng et al. [66] innovated with a sketching interface, empowering users to extract motion data by illustrating motion trajectories directly onto a screen. This method transcends the static nature of pose drawings [62], offering a dynamic canvas for envisioning character motion progression. However, the utility of such retrieved data is not fully realized without subsequent editing to align with specific user demands.

The Motion Doodle [81] framework tackled this by decomposing the sketched

motion trajectory into primitives, such as 'walk' or 'jump', and stitching these into coherent motion sequences like walk $\rightarrow$ jump $\rightarrow$ jump. Yet, its focus remains on 'global' motion, rooted in the character's hip movements, and thus, it does not adequately address the 'local' motion editing of limbs. In contrast, the SketchiMo system by Choi et al. [82] allows for direct manipulation of motion trajectories for specific body parts. While such systems enable meticulous editing of local motions, they also necessitate precise adjustments to the local coordinates of character joints at each frame, which can be intricate and time-consuming. MotionMaster [83], with its Kungfu motion database, introduces a methodology where users sketch initial and final poses and the trajectories of selected joint movements, yet the approach is marred by the requirement of refined drawing skills and laborious joint-pair labeling.

Advancing these methodologies, our framework proposes a bifurcated sketching process for motion trajectories: (i) the 'global' stage, concerned with the retrieval of root motions, and (ii) the 'local' stage, focused on the retrieval of limb movements. Through this division, we synthesize motion sequences by amalgamating the globally and locally retrieved motions, offering a more streamlined and user-friendly approach to character animation.

3.3 Global-to-Local Motion Design

The proposed character animation retrieval and composition scheme was inspired by the following observations: character animation design is challenging for novice users because of the complicated data format of motion data. In Addition, limited by their experience and skills, novice users may find the design task difficult as a single process and suffer from the blank page syndrome. To address this issue, we propose a two-stage design scheme to assist novice users in designing motions of character animations with both global and local stages. We decompose the whole motion design process into two stages. In each stage, users can just concentrate on one sub-goal of the design process. Users begin from the global stage and sketch the rough movement of the character's lower limb. After users are satisfied with the work, they enter the local stage and draw the detailed movements of the character's upper limbs. In this section, we go through the proposed motion design scheme and describe the global and local stages in detail.

3.3.1 Overview

In the proposed character animation authoring scheme, the character is shown on a virtual ground. The users draw query strokes that represent the motion trajectories of the character nodes. To fully utilize and reuse existing Mocap databases, we

conduct a trajectory similarity comparison between the user-input query strokes and the node trajectories in the database by projecting them into the same 2D camera coordinate system. Because the main purpose of the current prototype is locomotion retrieval, we only adopted the motions of walking, running, jumping, and punching and kicking for simplification. The overall pipeline is shown in Figure 3.2.

3.3.2 Global Stage (Lower Limb)

The global stage is the first stage and the default state of the two-stage design scheme. When users enter the stage by opening an initial scene, a blank virtual ground is displayed. To start the query and further composition, the users draw a trajectory on the screen, to represent the ground-projected moving trajectory of the character's center of mass. The system then starts to match the user-input query stroke with the projected trajectories in the database and apply the best-matched animation on the character in the interface. The users can choose the desired rough movement of the whole character and enter the local stage to specify more details of the character's upper limbs, or just repeat the retrieval in the global stage. In the global stage, we define the virtual ground as a reference point, which allows the users to freely adjust the position and angle of the camera.

3.3.3 Local Stage (Upper Limb)

After the users are satisfied with the retrieved global movement of the character, they can continue the design process and choose a specific upper limb for further detailed retrieval. The reference point of the local stage is the character's center of mass, which enables a surveillance camera that always follows the character and stare at the character's center of mass. The camera is located on a sphere whose radius is adjustable and the center is also the character's center of mass. Similarly, the users draw trajectories of the selected limb node with respect to the center of mass. When the users finish the query stroke input, the system translates all trajectories of the selected limb node from the database by projecting them to the camera coordinate system and starts similarity matching. The users then pick the desired limb motion from the retrieved results, and the system applies the motion to the selected limb based on the rough motion retrieved in the global stage to make a combined new animation. With iterations between global and local stages on different limbs, the users finally make a new animation from existing motion captures and save it to end the design process. Although legs should be considered as limbs, we only enable head and hand motion retrieval and composition in the local stage because leg motion is highly entangled with the overall movements of the character.

3.4 System Framework

In this section, we explain how we implement the motion editing system prototype called *DualMotion* in the proposed framework.

3.4.1 Dataset Construction

To build the data-driven motion editing system, we construct the motion dataset by selecting the locomotion data from the CMU Mocap data library [60], which consist of several motion categories, such as locomotion, physical activities & Sports, etc. (Note that our data is mainly chosen from the locomotion.) We empirically choose motion data with similar sizes and skeletal proportions. We then manually trim the motion data into 100 frames to allow them to share the same duration, and normalize all the motion sequences into the same initial position that is the origin of the virtual spatial coordinate system. In our case, the root node (hip) movement of each motion data is considered the global movement, and each limb movement is considered the relative local movements of the root node. Lastly, each motion data point is split into limbs and hips, which are stored.

3.4.2 Trajectory Representations

The users directly sketch on the widget that visualizes the character animation for both global and local stages. That is, the sketching canvas coordinate system shares with the camera coordinate system of the visualization widget. Note that as stated in Section 3.3, the editing tool maintains a camera with different reference points in the global and local stages.

In each stage, motion trajectories in the motion dataset are projected into the canvas coordinate system, as follows:

$$\mathbf{V}^j_{canvas}(i,t) = M_{proj}\mathbf{V}^j_{orig}(i,t) \tag{3.1}$$

where $\mathbf{V}^j_{orig}(i,t) \in \mathbb{R}^3$ is the 3D coordinate of i-th node at time $t \in \{1,T\}$ in the j-th motion data in the database, $\mathbf{V}^j_{canvas}(i,t) \in \mathbb{R}^2$ is the corresponding projected 2D coordinates of the motion trajectory, and $M_{proj} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is a 3D to 2D projection matrix of the maintained camera. In this paper, we represent the motion trajectory in the database as follows:

$$V^j_{orig}(i) = \{\mathbf{V}^j_{orig}(i,1), \mathbf{V}^j_{orig}(i,2), \cdots, \mathbf{V}^j_{orig}(i,T)\} \tag{3.2}$$

where T is the number of the motion frame. When the users pan or zoom to relocate the camera so that the projection matrix M_{proj} changes, the proposed tool updates the projected trajectories $V^j_{canvas}(i) = \{\mathbf{V}^j_{canvas}(i,1), \cdots, \mathbf{V}^j_{canvas}(i,T)\}$ for comparing the user query sketch with the motion in database.

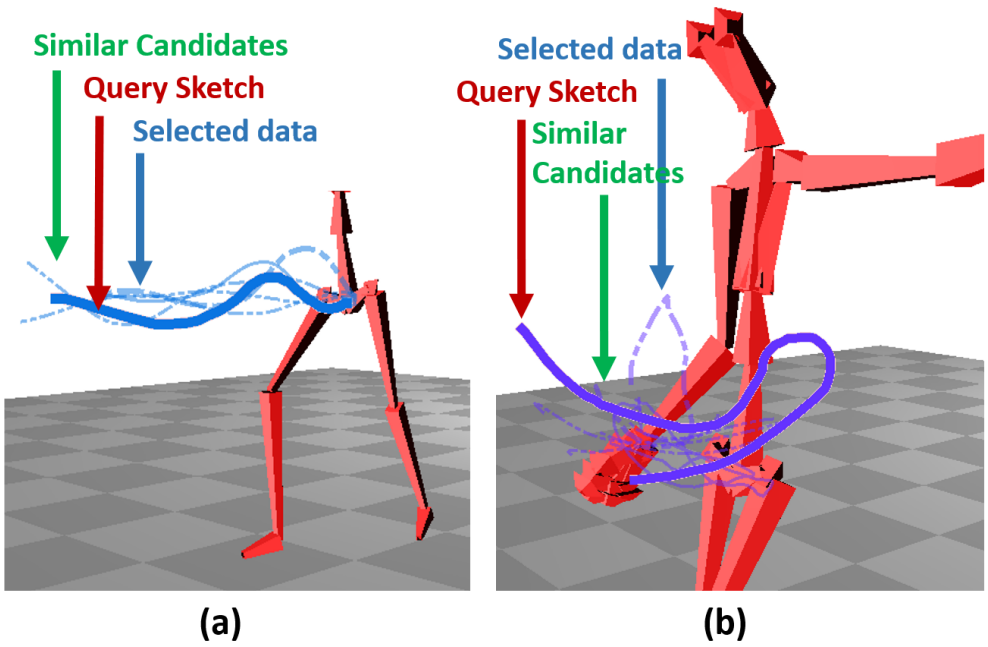

Figure 3.3: Shadow-like guidance that displays joint movement projected trajectory in global (left) and local (right) design stages.

3.4.3 Trajectory-Based Retrieval

Similarly, we represent a user-input stroke S_{user} for both two stages as follows:

$$S_{user} = \{\mathbf{S}_{user}(1), \mathbf{S}_{user}(2), \cdots, \mathbf{S}_{user}(T)\} \tag{3.3}$$

where $S_{user}(t) \in \mathbb{R}^2$ is the coordinate of t-th sampling point of the user-input stroke, and T is the number of total sampling points of the stroke.

We then compute a similarity between the query stroke S_{user} and the projected trajectory of the j-th motion data in the database V^j_{canvas}, as follows:

$$Sim^j(i) = F(S_{user}, V^j_{canvas}(i)) \tag{3.4}$$

where $F(a, b)$ is the Fréchet distance between two strokes a and b. Note that i represents the root node in the global stage, and the limb nodes in the local stage.

After computing all the similarities, we pick the top N relevant retrieval results in the database ($N = 5$ in our implementation) and merge them as shadow-like guidance to display their projected trajectories in real-time, inspired by *ShadowDraw* [84]. As shown in Figure 3.3, the user can click on the most desirable motion candidate for selecting global/local motion data. Benefiting from the guidance, users understand the locations and shapes of each stroke drawing.

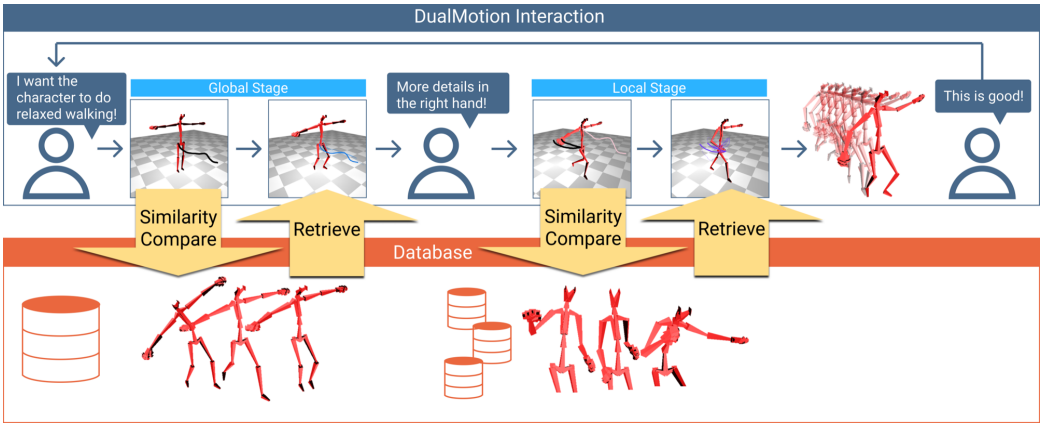

Figure 3.4: The design process of *DualMotion*.

3.4.4 Motion Sequence Synthesis

Given the retrieved global and local motions, *DualMotion* generates a final motion sequence. In our prototype, we utilized a BVH data format, which consists of the position of the root node and all joints' rotation at each frame, and simply assigned the rotational matrices of the associated nodes in the retrieved local motion to the retrieved global motion. For instance, the rotational matrices of the left shoulder node and its child nodes would be modified when the left-hand movement is authored. Similarly, the right shoulder node, lower neck node, and their child nodes are allowed to be re-assigned when the related movement is modified. A similar approach is applied by Iwamoto et al. [79] who segmented the hand movement and body motion to synthesize the dance motion sequence. Although this approach seems to work well, other methods can also be used in our framework.

3.4.5 User Interface

We implement the two-stage character animation design scheme as a prototype called *DualMotion* using OpenGL and Qt. Figure 3.6 shows the user interface of *DualMotion*. It consists of a graphical widget and a control panel. The graphical widget is the main region for users to interact (visualization and drawing). At the top of the graphical widget, the control panel provides buttons to enable basic functions (undo, load, save, etc.), mode selection between drawing and result selecting, camera controlling, global and local stage switching, and timeline visualization. We have introduced the basic design scheme in Section 2.3. As an alternative for showing the character animation shown in Figure 3.5, we implemented a timeline

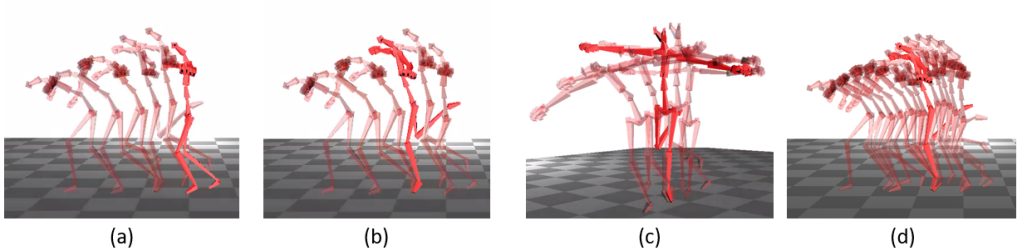

Figure 3.5: (a) The initial state of action sequence visualization, (b) selection of action data for a specific frame, (c) free change of viewpoint to observe the data, (d) interpolation to change the number of visualized selected frames.

visualization to help users check the entire animation in a static way rather than waiting for the animation playback. The character animation is shown in the visualization widget by sampling the frames in the motion sequence, and the character skeleton of each frame is statically overlaid in the graphical widget. Only one frame of the character skeleton is highlighted, whereas the others are semi-transparent. A slider is used for adjusting the position of the highlighted frame, and another slider is used for manipulating the frame interval. In this way, users are not only able to search for and edit skeleton animations by motion trajectories, but they can also freely change the viewpoint and frame rate with *DualMotion* to observe and select poses for individual frames. Moreover, since the poses selected from the motion data in the search database originate from real data, such poses are more vivid and natural compared to those from freehand sketches, and will not contain artifacts.

3.4.6 Design Process

The design process of *DualMotion* is shown in Figure 3.4. The user starts from the global stage with a rough objective about the character animation in mind and then goes forward into the local stage and decides on more details about the upper limbs. The user can repeat the two design stages until satisfied with the final output. Every time the user inputs query strokes, the *DualMotion* interface feeds forward the user inputs, compares them with the database and renders the retrieved and composited motion sequences.

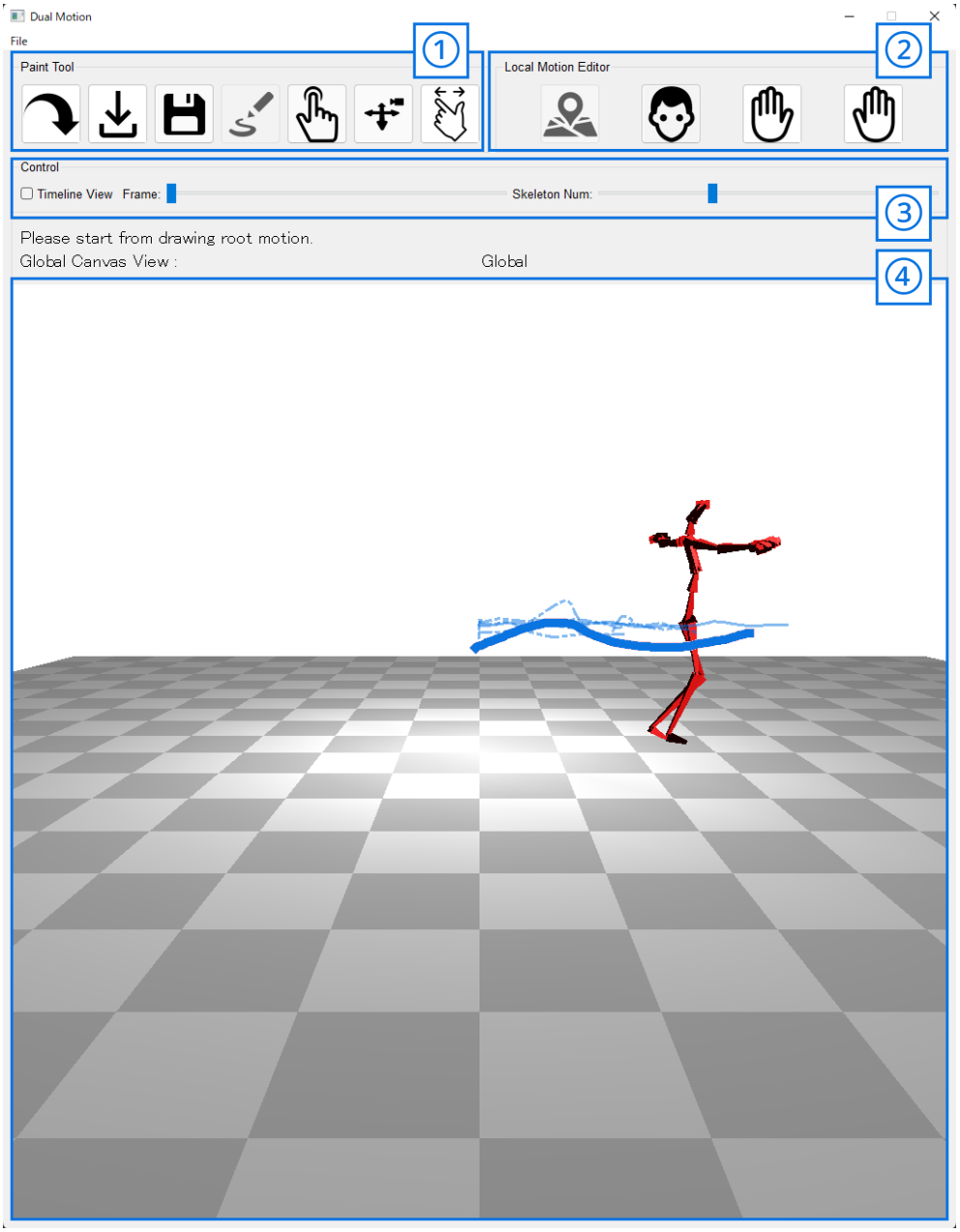

Figure 3.6: Interface of *DualMotion*. It consists of: ① a tool panel contains buttons like undo, load, save, draw, select, camera pan, and camera zooming; ② a stage panel to switch between global and local stages; ③ a control panel for timeline visualization and ④ a graphical interface for drawing and character animation visualization.

3.5 User Study

To evaluate the validation of *DualMotion*, we conducted a user study that consists of a comparison study and a post-experiment questionnaire. In this section, we describe the experimental details and discuss the evaluation result.

3.5.1 Comparison Study

To evaluate the validation of the proposed user interface, we conduct a user study to compare *DualMotion* with a one-stage motion retrieval UI, which has the same function as *DualMotion*, except that it only allows users to edit and synthesize motion to obtain the results but only in the global stage editing (i.e. hip movement). For example, users are allowed to edit a hand-wave movement while the character is stepping forward. It is difficult for users to define such a hand movement on a moving character in this case. Similar to *DualMotion*, the implemented one-stage UI provides shadow guidance to aid users in finding similar results (trajectories of motion) after each time they finish drawing a stroke. However, in contrast to *DualMotion*, it only allows users to retrieve motion by drawing the trajectories in a global manner.

The study consists of two sections: 1) motion design with a reference motion, and 2) free motion design. In the first section, we provided five skeleton motions that consist of different identities of joint movements from the prepared database. By using *DualMotion* and the one-stage edit UI respectively, the participants were required to retrieve and edit the motion following one of these references randomly.

On the one hand, to compare the efficiency, we recorded the time cost and the number of operations (mouse clicks). On the other hand, to compare the quality of the design result, we also calculated the mean absolute error between the synthetic motion and the reference motion to compare the similarity, which confirms the accuracy of the user design result in the two different UIs. In the second section, to evaluate whether our interface meets user intent, the participants are asked to do a free motion design using *DualMotion* in 5 minutes. (Our own trials with the system revealed that by repeatedly drawing the motion trajectories to search and edit actions to match our desired outcome, the entire process generally does not exceed 5 minutes.) They are asked to answer their editorial intent or the reason why they would like to design such kind of a character motion in the post-experiment questionnaire.

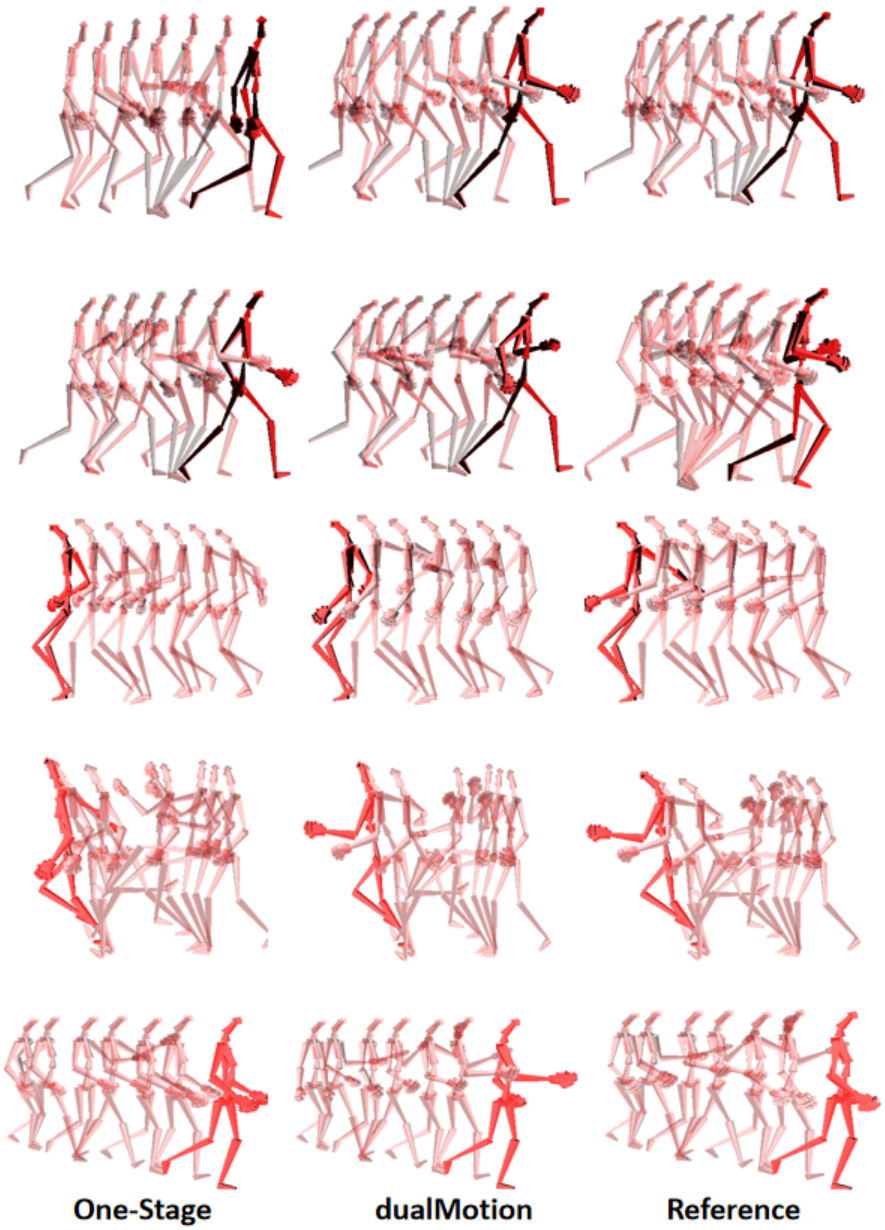

Figure 3.7: Examples of participant design results by using each interface for the five motion references. The design results using *DualMotion* are more similar to the reference compared to the one-stage interface, especially the hand movements.

3.5.2 System Evaluation

Except to answer the editorial intent in the free design task, the participants are required to evaluate the overall system by answering the questions relying on System Usability Scale (SUS) [85] metrics after the experiment. To rate the perceived workload of design motion using *DualMotion*, we requested the participants to fill in a post-experiment questionnaire designed following NASA Task Load Index [86] (Considering the length of the questionnaire, we utilized the *raw* version of NASA-TLX in our evaluation study where weighing each of the evaluation items is not necessary).

3.5.3 Experiment Process

We invited 14 graduate students at the age of approximately 25 years to participate in our experiment. We choose a total of 55 skeleton motion data from the CMU [60] motion database for retrieval in the user study, which are gait data in different directions with different styles, including fast walk, slow walk, duck walk, zombie walk, and so on.

They had approximately 5 minutes to get familiar with each tool before the task. In addition, a player window with a modifiable virtual camera is provided so that reference motion clips can be viewed at any time during design. The order of finishing tasks using one-stage UI and *DualMotion* are randomly shuffled to make sure that we equally evaluate each UI. The free motion design experiment is the final task that we believed that the participants are familiar enough to user *DualMotion* to create some motion following their intent. Lastly, they are required to fill out the post-experiment questionnaire mentioned above.

3.6 Results

3.6.1 Evaluation Results

Figure 3.7 (right) shows various reference motions used in our user study. For the validation of *DualMotion*, we compared the Euler angle of i-th node at time $t \in \{1, T\}$ in the user-designed results $\mathbf{x}_i(t)$ to the reference $\mathbf{x}'_i(t)$ and computed the MSE loss as follows:

$$MSE = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \|\mathbf{x}_i(t) - \mathbf{x}'_i(t)\|^2 \tag{3.5}$$

where N is the number of joints in the character skeleton. The edited motion results by using *DualMotion* have a higher similarity compared with the one-stage

design approach. In addition, we analyzed the overall MSE result of *DualMotion* by running a one-tailed t-test. The t-value was 2.48, and the p-value was 0.01 ($p < 0.05$), which revealed significant differences in the MSE loss.

For the evaluation of the efficiency, Figure 3.9 shows the average time cost and operation times (in this case, we recorded the number of mouse clicking during the task). It illustrates that *DualMotion* allowed users to design motions more efficiently. Notably, the time cost and the operation times of the free motion design task (Figure 3.9 (right-most boxes)) included the time for envisioning their own target motion for each participant. Therefore, the participants are able to manage *DualMotion* and create the motion following their intent in a short time.

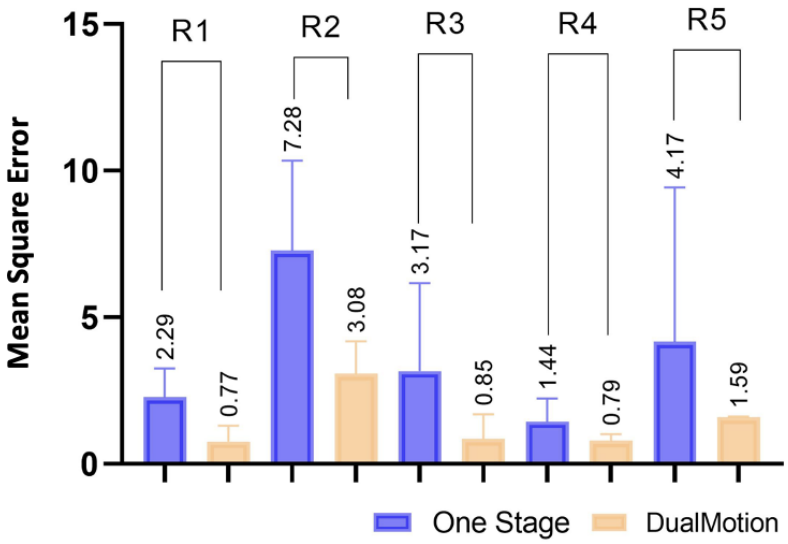

Figure 3.8: MSE loss results between the user design results and the references. (R1, R2, etc are indexes of the reference motion.)

We show the post-experiment questionnaire results below. The SUS metrics evaluation result is shown in Table 3.1. All of the participants claimed that they felt satisfied with the overall interface, and they imagined that most of the people would learn to use this system very quickly. Furthermore, 86% (12/14) of the participants said that they would like to use *DualMotion* frequently to design the prototype of character animations. Moreover, 79% (11/14) and 71% (10/14) of the participants said that they were confident when using the system and that the functions were well-integrated, respectively. *DualMotion* scored 75.6 out of 100 through the SUS metrics, which implies overall good usability. (based on various studies [87–89], the average SUS score is around 68-70.5. Therefore the SUS score of the proposed system was beyond the average.)

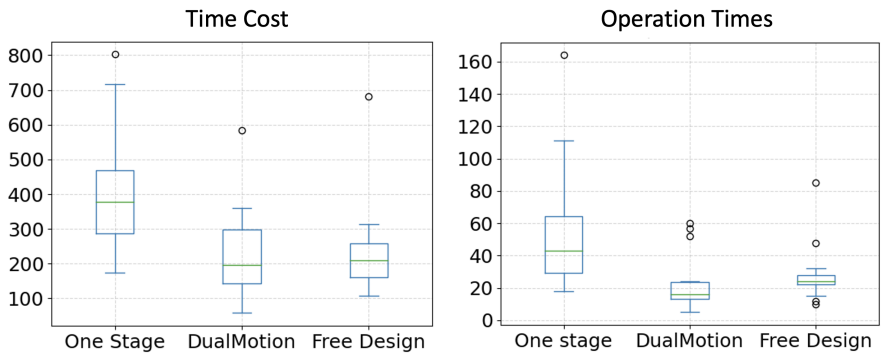

Figure 3.9: Average time cost (left) and operation times (right) of each task. Note that the free design task was only done by using *DualMotion*. It shows that the participants were able to design an expected motion within a few minutes.

Lastly, the workload evaluation result is shown in Figure 3.10. The high score of '*Overall Performance*' illustrates that the participants were satisfied with their own animations. By contrast, the low level of '*physical demand*' and '*frustration level*' implies that the large loads of labor are unnecessary when utilizing *DualMotion* to design a character animation.

3.6.2 Animation Prototype

To evaluate whether the design results are qualified enough to make an animation prototype, we retargeted the users' free-designed skeletal motion onto several character models with various body sizes (Michelle, Maynord, Ninja, and Ortiz from *Mixamo* [90]). We observed that the animation prototype was consistent with the user's editorial intent. As shown in Figure 3.11, the user editorial intents were *a crash* (a) *a zombie walk* (b), *a standing long jump* (c), and *a run with joy* (d). More results are shown in the uploaded video https://www.youtube.com/watch?v=-tk8q8LSiL0. We also conduct oral interviews with users after finishing their free-designed animation. Most of the users said that although they only had a rough idea at the very first and did not clearly know exactly how to implement movement (e.g., someone wanted to design a zombie walk but did not exactly know how the hands or head move.), they can achieve their goal and were satisfied with the results. The results verify that *DualMotion* can not only support high-precision motion design but can also help users successfully complete their animations with only vague ideas and achieve satisfaction with their own work.

Table 3.1: Results of the post-experiment SUS metrics questionnaire. ⇑ indicates higher scores are better. ⇓ for the other case. The total score is 75.6 out of 100.

#	Questions	Mean	SD
1	I think that I would like to use *DualMotion* frequently. ⇑	4.00	0.48
2	I found *DualMotion* unnecessarily complex. ⇓	2.07	0.92
3	I thought *DualMotion* was easy to use. ⇑	3.93	0.99
4	I think that I would need the support of a technical person to be able to use this *DualMotion*. ⇓	2.64	1.28
5	I found the various functions in this *DualMotion* were well integrated. ⇑	3.86	0.66
6	I thought there was too much inconsistency in *DualMotion*. ⇓	1.92	0.91
7	I would imagine that most people would learn to use *DualMotion* very quickly. ⇑	4.50	0.52
8	I found *DualMotion* very cumbersome to use. ⇓	2.07	0.99
9	I felt very confident using *DualMotion*. ⇑	4.07	0.73
10	I needed to learn a lot of things before I could get going with *DualMotion*. ⇓	1.42	0.51

3.7 Discussion and Limitation

3.7.1 Applications on Mobile Platforms

In our user study, we found that *DualMotion* supports novices to create character animations in two usage scenarios: 1) to create a character animation with a clear target (given a reference motion sequence), and 2) to create a character animation without a target. Particularly, *DualMotion* reduces the operations in casual character animating. As an extreme case, animating on mobile devices with touch screens strictly limits the count and accuracy of user operation. Actually verifying *DualMotion*'s extendibility on tablets and smartphones is one of our future work.

3.7.2 Trade-off between Dataset Size and Computation Performance

The current prototype of *DualMotion* and the user study are based on a lightweight motion database. (55 motion data in total) Our system updates the projection matrix every time the camera is relocated by users and then applies to all motion

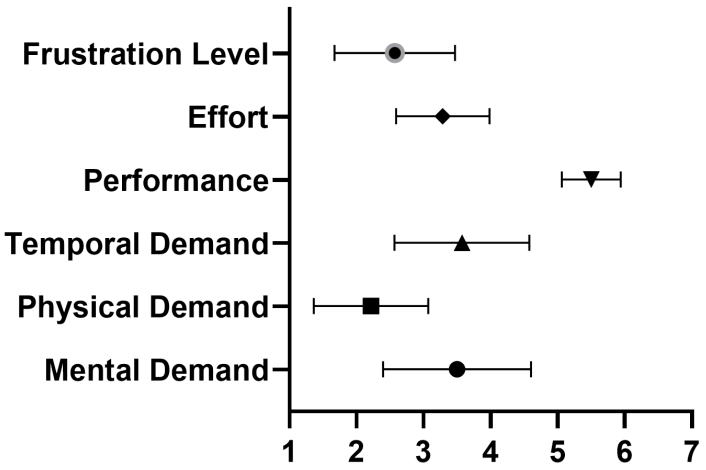

Figure 3.10: The result of *raw* NASA-TLX evaluation.

sequences in the database. The average retrieval time is around 0.84 seconds for each query. As the size of the database grows, the computation performance may not be able to meet the demands of real-time interaction sensitively. It mentioned that the retrieval task is implemented by a single node loop on the CPU, and a parallel processing pipeline on the GPU is required to increase the efficiency. Besides, we will extend the current database and accelerate the algorithm to improve the diversity of design results.

3.7.3 User Sketches in Design Process

The current user-input stroke is made up of uniformly distributed points, which means that the input omits the velocity of the user drawing. We only enable head and hand node authoring in the prototype for simplicity of the design process; however, it limits the diversity of design results. Exploration of inputs that contain more information is yet another future work for more professional users.

3.7.4 Creativity Support with Two-Stage Scheme

DualMotion adopted a two-stage scheme with both global and local stages [58, 91]. Along with the perception capabilities of humans, designing the target with a bottom-up scheme is difficult. For example, we usually have a gradually clear design intention during the design process and only an ambiguous target in our mind at the initial step. We thought the two-stage scheme is an approach to human-centered design thinking and could be useful in various creativity support systems, such as illustration and story designs.

As technical limitations, the current prototype of *DualMotion* combined the limbs' motion data with the torso's of the character model. The current implementation may produce artifacts such as unnatural body balance. A lightweight algorithm for automatically correcting the full body balance can be considered for further development, such as a double inverted pendulum model [92].

3.8 Conclusion

In this work, we proposed *DualMotion*, a casual motion design UI for character animations. This helps common users create character motion animation, as they are not always able to conduct accurate and large amounts of operations. In this case, editing motion by using professional software, which has many complex functions when the operation count and range are limited, poses further challenges. The proposed system utilizes a two-stage design scheme with global and local motion retrieval and composition to tackle the challenging task of decomposing full-body motions into lower-limb and upper-limb movements. Users are allowed to query the database with simple and rough sketch input and retrieve desirable results in a data-driven manner. Besides, *DualMotion* offers a considerable speed query for motion data retrieval. For instance, the number of keypoints in a stroke will be comparatively fewer when a motion trajectory is drawn rapidly, resulting in a relatively faster motion retrieval outcome.

In addition to retrieval and editing motion data, *DualMotion* can produce a series of related action sequences from vague inputs such as motion trajectories, providing users with a selection of specific poses for individual frames. This is similar to a photographer taking several random shots of a model in various poses and angles before selecting the photo that best fits their vision. It also offers effective suggestions for specifying single-frame poses.

Lastly, a user study was conducted, which verified that *DualMotion* is available to support users to create character animations in low labor demand, even with only vague ideas. Based on the analysis of user feedback from our user study, most users acknowledged the ease of use of *DualMotion*. However, some users pointed out the difficulty in editing long motion sequences with *DualMotion*. This challenge arises because the motion trajectories in long sequences are more complex, making it inconvenient for users to specify overly intricate trajectories on a fixed-view 2D drawing board. Therefore, we envision that editing long motion sequences should ideally be based on specifying and combining multiple short sequence actions. We plan to implement and refine this functionality in our subsequent work.

In summary, *DualMotion* demonstrates a high capability for understanding vague inputs (Flexibility), along with a strong ability to generate diverse outcomes

(Variability). It also maintains a lower level of one-to-one Precision between input (motion trajectories) and output (motion data) to ensure the diversity of retrieval results.

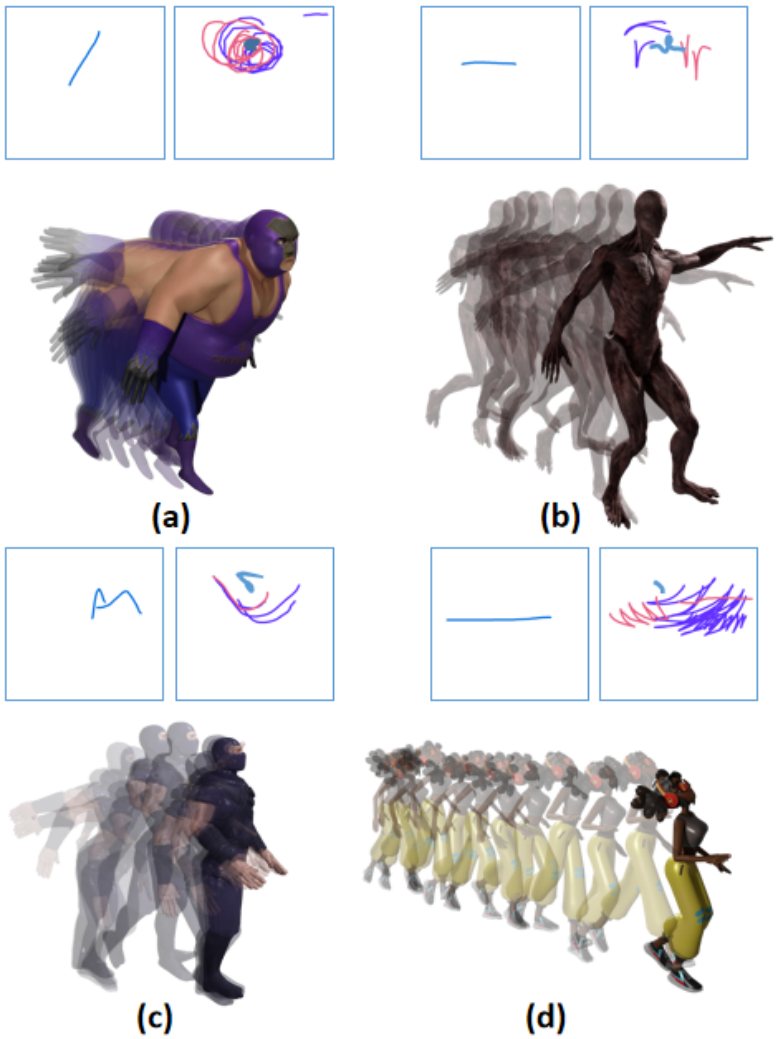

Figure 3.11: Motion retargeting results on Mixamo's character models. The query sketch is shown above each result which are global editorial strokes (left) and local editorial strokes (right). The different colors in the local design stage represent the design history of various joints which are the head (cyan), left hand (violet), and right hand (pink) respectively.

Chapter 4

Hierarchy Style Injection for Human Image Generation via Diffusion Model

In the preceding chapter, we introduced a method that ingests motion trajectories to search and synthesize animations, offering a novel and effective approach for users to design character animations and specify poses from multiple perspectives. In this chapter, we propose a generation method based on the pretrained diffusion model fine-tuning, which facilitates the creation of human images from specified poses through the hierarchy style injection. This approach enables precise control over the image generation process, accommodating user requirements for pose generation in human-centric scenarios.

4.1 Introduction

The task of pose and style-guided human image synthesis has consistently been a focal point of exploration in generative models. This technique holds potential for widespread applications across various domains including character design, apparel creation, and the generation of diverse content within digital realms, with implications for both commercial and entertainment industries. Moreover, recent studies have leveraged these methodologies to bolster performance in downstream tasks such as merchandise recognition, specifically in clothing, and individual identification. These applications underscore the necessity for models to exhibit robust capabilities, such as the decoupling of pose and style attributes and the generation of fine details.

Generative modeling has seen a surge in innovation with the rise of applications necessitating the synthesis of human images that are both realistic and style-consistent. In this realm, pose and style-guided image generation stands out as a particularly challenging yet impactful area of study. The ability to generate images of humans in arbitrary poses with specific styles has implications extending beyond mere visual synthesis; it has the potential to revolutionize industries by providing a foundation for advanced character and fashion design, enhancing digital content creation, and offering improvements in visual media production.

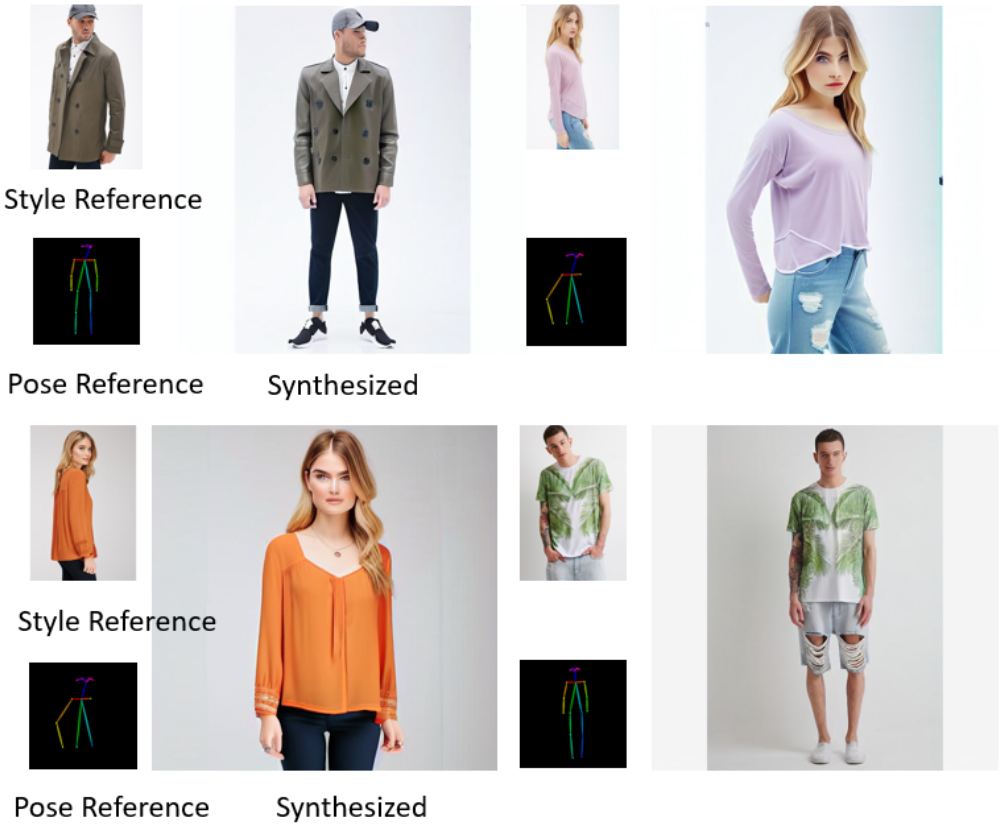

Figure 4.1: Pose & Style guided person image synthesis. Our model enables users to control the generation of human images by inputting the desired pose and a reference clothing style. It ensures that the generated results exhibit high consistency with the features of the input clothing and the specified pose.

However, the challenge of person synthesis is typically addressed using Generative Adversarial Networks (GANs) [14], which attempt to generate a person in a desired pose through a single forward pass. Yet, maintaining coherent structure, appearance, and overall body composition in a new pose remains a formidable one-shot task. The outputs often feature deformed textures and unrealistic body forms, particularly when recreating occluded body parts. Additionally, due to the adversarial min-max objective, GANs are susceptible to unstable training behaviors, resulting in a lack of diversity in the generated samples. In contrast, solutions based on Variational Autoencoders [93] are more stable but tend to produce outputs with blurriness and lower quality compared to GANs, stemming from their reliance on a surrogate loss for optimization.

In this work, we propose a fine-tuning approach for the large-scale pre-trained text-to-image diffusion model, HSI, and have trained a pose & style-to-human generative model. This enables users to generate personal images by freely altering poses and styles. Furthermore, our model facilitates editing and the trial of various ensembles and combinations. By leveraging the robust foundation of the HSI model, our approach not only enhances the flexibility of pose and style synthesis but also significantly improves the user's creative agency in the generative process. The capability to iterate over different configurations allows for a tailored and more personalized output, reflecting the user's specific aesthetic and functional requirements.

Our contribution can be summed up as follows:

- HSI architecture is proposed for consistent style sampling. This architecture utilizes selected hidden states from the pre-trained CLIP vision encoder, which are injected into the U-Net's cross-attention units to maintain style fidelity across generated images.
- A Covariance-based Pose and Style Decoupling (CoPSD) method is introduced to disentangle the fusion of input pose and style reference biases. This is achieved by optimizing a covariance loss between pose features and style features during fine-tuning, thereby enhancing the model's ability to independently manipulate pose and style.
- The Cross-Pose Style Integration Training (CPSIT) strategy is developed to augment the adapter module's robustness. It involves unpairing the training data during the fine-tuning stage to enable the model to handle a wider variety of poses and styles, improving its generalization capabilities.

4.2 Related Work

4.2.1 Conditional Generative Models

In Chapter 2.1, we review a variety of generative models and algorithms, with conditional generative models, especially cGANs, being the most readily utilized for downstream tasks. However, when input conditions involve multiple features, such as pose and style in this work, the generative model requires a robust feature decoupling capability to alter one input without affecting the others. StyleGAN and its various derivatives, such as pixel2style2pixel [94], are recognized for their excellent feature decoupling abilities.

As mentioned in chapter 4.1, especially in tasks that demand high feature decoupling, GAN-based models encounter stringent training conditions and are prone to issues such as mode collapse. On the other hand, generative models based on text-to-image diffusion, such as Stable Diffusion [24] and GLIDE [20],

to some extent circumvent these problems. While these models exhibit powerful performance, training such large-scale generative models requires extensive data and computational power. Consequently, many recent works focus on fine-tuning these large models, examples being ControlNet [26] and T2I-adapter [27].

4.2.2 Pose-Guided Human Image Synthetic

The task of pose-guided person image synthesis is regarded as a form of exemplar-based image translation, where the objective is to replicate the appearance from reference images under arbitrary poses. Early approaches [93, 95] addressed this issue by extracting pose-irrelevant vectors to encapsulate appearance. Nonetheless, considering the vast variation in textures across different semantic entities, directly deriving vectors from reference images may restrict the model's ability to represent complex textures. Additionally, models such as text2human [96] and UPGPT [97] offer users the capability to generate and edit human images through text-based control of specific segmentation regions, enabling targeted input manipulation. However, these tasks initiate training from scratch, necessitating a substantial amount of computational resources.

In this work, we introduce a fine-tuning methodology for the HSI, a large-scale pre-trained text-to-image diffusion model, and have developed a pose & style-to-human generative model. This model empowers users to create personalized images by freely modifying poses and styles. Moreover, it supports the editing process and experimentation with a range of ensembles and combinations.

4.3 Method

In this section, we first delineate the preliminary knowledge of attention mechanisms within diffusion models for conditional generation tasks. Following this foundational overview, we expound upon the motivations driving this work and provide a detailed account of the methodologies employed in our approach.

4.3.1 Preliminary for Diffusion Model

Diffusion models, a novel class of generative models, operate through a process that can be divided into two phases: a noise-adding diffusion process and a noise-reducing sampling generation process. The diffusion process incrementally adds Gaussian noise to the data in a manner that follows a predefined Markov chain, with a fixed number of diffusion steps, T. The denoising sampling process, on the other hand, involves training a learnable model (typically a U-Net architecture) to take in the noised data X_t at a given time step t, along with the timestep T itself,

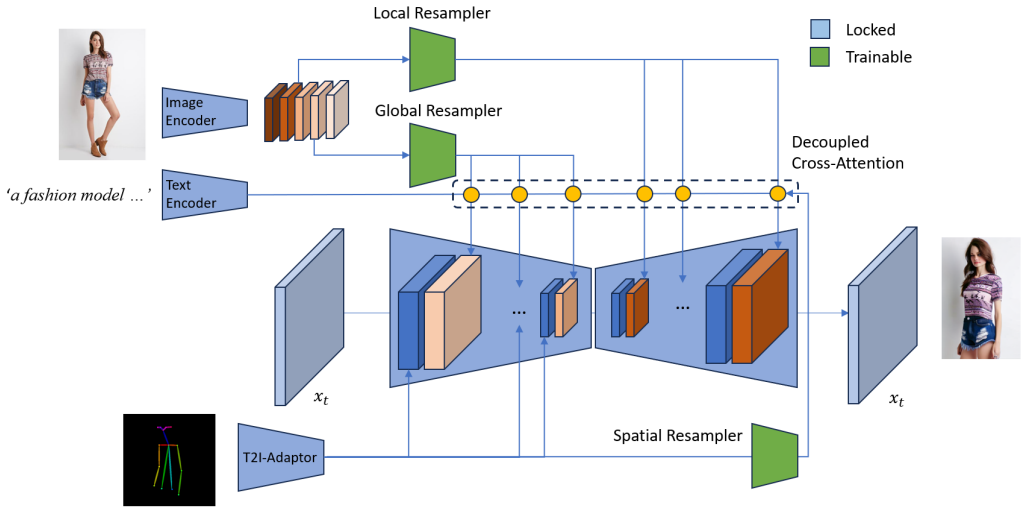

Figure 4.2: Hierarchy Style Injection (HSI) and Covariance-based Pose and Style Decoupling (CoPSD). This schematic demonstrates the integrated HSI-CoPSD framework for style control and pose preservation. HSI leverages a pretrained encoder' s intermediate features for style modulation via cross-attention in a denoising U-Net, while CoPSD ensures independent pose and style feature manipulation, mitigating the risk of pose distortion due to style variance.

and predict the noise that was added at the previous step $t-1$ in order to reverse it. Currently, diffusion models are frequently applied in conditional generation tasks such as text-to-image synthesis, where the text condition is denoted as c. The training objective of a diffusion model, denoted as ϵ_θ, which is tasked with predicting the noise, is often defined as a simplified variant of the variational lower bound:

$$L_{Simple} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(x_t, c, t)\|^2] \tag{4.1}$$

To formalize, in diffusion-based generative tasks, x_0 represents the clean, real data subjected to an additional condition c. The diffusion process is indexed by time steps $t \in [0, T]$, where $x_t = \alpha_t x_0 + \sigma_t \epsilon$ denotes the noisy data at the t-th step. The functions α_t, and σ_t dictate the noise level, while ϵ is a noise vector.

Upon completion of training, the model ϵ_θ is adept at predicting the noise introduced at each step. The generation of images post-training is an iterative process, beginning with random noise. The model progressively reverses the noise addition process, at each step being guided by the condition c, culminating in the reconstruction of the clean data x_0.

In the realm of conditional diffusion models, classifier guidance is often em-

ployed to enhance image fidelity while maintaining sample diversity. This is achieved by leveraging the gradients from a separately trained classifier [23]. To obviate the need for an independent classifier, classifier-free guidance presents an efficacious alternative [39]. It entails the simultaneous training of conditional and unconditional diffusion models with the condition c being randomly omitted during the training process. During the sampling phase, the noise is predicted by considering the outputs from both the conditional model $\epsilon_\theta(x_t, c, t)$ and the unconditional model $\epsilon_\theta(x_t, t)$:

$$\epsilon_\theta(x_t, c, t) = w \cdot \epsilon_\theta(x_t, c, t) + (1 - w) \cdot \epsilon_\theta(x_t, t), \tag{4.2}$$

Here, w, commonly referred to as the guidance scale or guidance weight, is a scalar that modulates the model's adherence to the condition c. In text-to-image diffusion models, classifier-free guidance is pivotal for augmenting the congruence between the generated images and the textual descriptions they are conditioned upon.

4.3.2 Hierarchy Style Injection (HSI)

In the realm of traditional text-image diffusion models, the "conditioning" input is typically synonymous with textual input. Innovations such as ControlNet, T2I-adapter, IP-adapter, and MasaCtrl build upon the original diffusion model framework. These systems incorporate externally trained modules designed to extend the diffusion model's compatibility with various modalities, including sketches, segmentation masks, depth maps, and poses. Notably, IP-adapter and MasaCtrl are efforts focused on modulating the sampling process by calibrating the denoising U-Net's cross-attention mechanisms. This optimization of the generative process is achieved by steering the cross-attention units within the denoising phase. A simplified exposition of the attention mechanism is delineated herein.

Inspired by the IP-adapter approach, this paper proposes HSI, a method that leverages the intermediate layer features of a pre-trained CLIP Vision Encoder. This hierarchical strategy involves the integration of auxiliary cross-attention modules into a denoising U-Net to control the noise-reduction sampling process. Our observations have led to the following insights:

1. As illustrated in Figure 4.3, the denoising U-Net, during sampling, exhibits a proclivity for global feature consideration in the downsampling stages while emphasizing local detail in the upsampling stages.
2. The pre-trained CLIP Vision Encoder we employ is a transformer-based architecture. Consequently, its outputs from the deeper layers, having undergone multiple convolutional operations, retain more high-dimensional abstract features. Conversely, the outputs from the earlier layers, subject

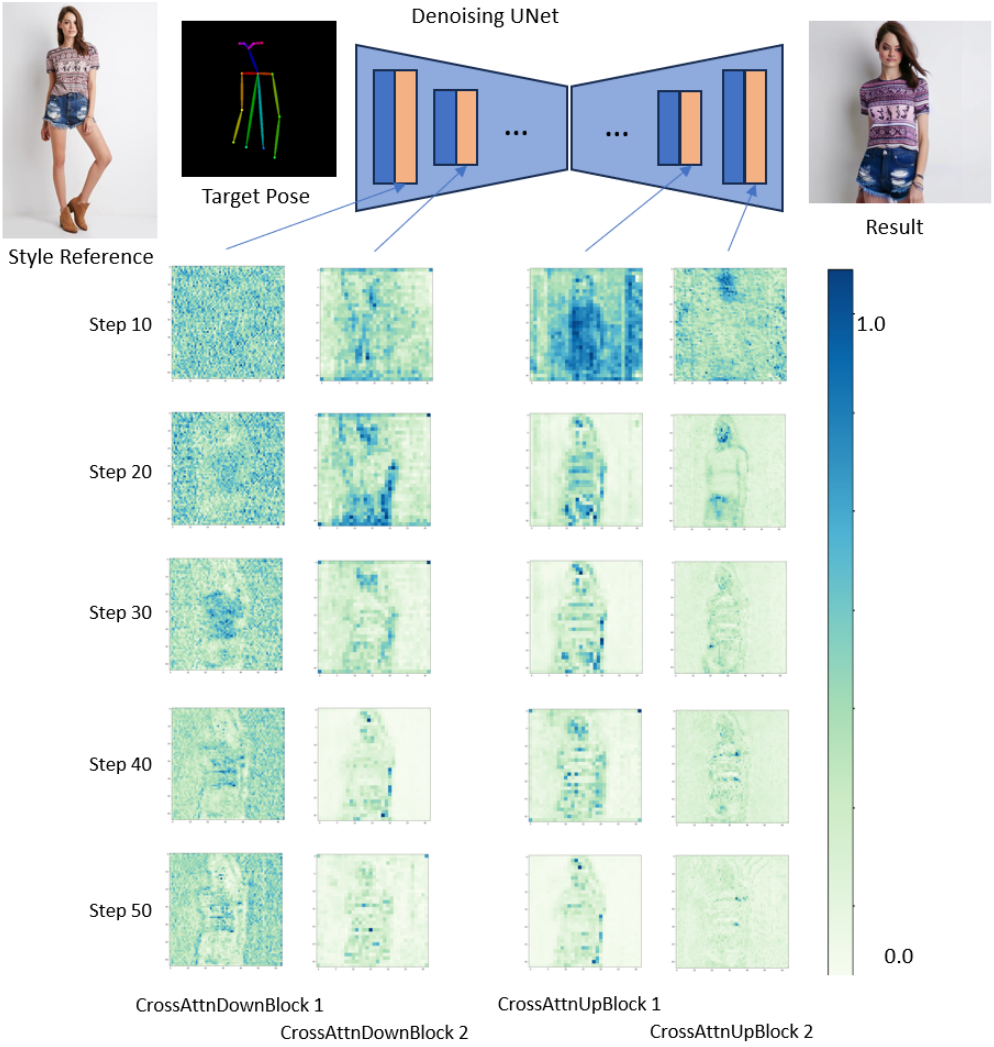

Figure 4.3: Denoising U-Net Sampling and Cross-Attention Maps on DDIM Schedule. This visualization showcases the outcomes at the 50th sampling step under a Denoising Diffusion Implicit Models (DDIM) schedule alongside intermediate cross-attention maps recorded at every 10th step. Note that intermediate layers with attention map resolutions at or below 16×16. are omitted for clarity and brevity in presentation.

to fewer convolutions, preserve more local detail. Figure 4.3 shows the encoding distribution of CLIP Vision Embedding across 10 sets of various types of clothing images. We find that, in comparison, the silhouette scores of the 2nd and 4th layers of the transformer perform better.

Drawing upon these insights, our proposed framework employs the second-to-last layer outputs of the CLIP encoder to train a global resampler. This resampler is designed to map features into the cross-attention layers of the U-Net associated with downsampling. Similarly, the fourth-to-last layer outputs are used to train a local resampler, aimed at mapping local detail into the U-Net's upsampling layers, as shown in Figure 4.2 This approach allows the diffusion model to more accurately reflect the stylistic nuances of the image prompt.

Originally, given the query features $\mathbf{Z}$ and the text features $\mathbf{c}_t$, the cross-attention output $\mathbf{Z}'$ is defined by the equation:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Z}, \mathbf{c}_t) = \text{softmax}\left(\frac{\mathbf{QZ} \times \mathbf{K}_c^\top}{\sqrt{d_k}}\right)\mathbf{V}_c \tag{4.3}$$

where $\mathbf{QZ}$, $\mathbf{K}_c$, and $\mathbf{V}_c$ represent the query, key, and value projections of $\mathbf{Z}$ and c_t respectively, and d_k denotes the dimensionality of the keys. With the introduction of distinct global and local feature descriptors, represented by $\mathbf{K}_{global}$, $\mathbf{K}_{local}$, $\mathbf{V}_{global}$, and $\mathbf{V}_{local}$, respectively. we redefine the updated feature vector Z_{new} as a function of these newly acquired key and value pairs by summing up the global & local attention with a weight w, written as:

$$\begin{aligned}\mathbf{Z_{new}}' = \text{softmax}\left(\frac{\mathbf{QZ} \times \mathbf{K}_c^\top}{\sqrt{d_k}}\right)\mathbf{V}_c + \\ w \times \left(\text{softmax}\left(\frac{\mathbf{QZ} \times \mathbf{K}_{global}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{global} + \right. \\ \left. \text{softmax}\left(\frac{\mathbf{QZ} \times \mathbf{K}_{local}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{local}\right)\end{aligned} \tag{4.4}$$

4.3.3 Covariance-based Pose and Style Decoupling (CoPSD)

During our experiments, we observed that the pose and style features within the dataset may be entangled, leading to inadvertent shifts in the output pose when altering only the style reference image under a constant pose condition. Furthermore, during the fine-tuning phase, the conventional loss function, denoted as L_{Simple}, tends to converge rapidly and ceases to decrease (as shown in Figure 4.5), indicating that using L_{Simple} alone is insufficient for further optimization

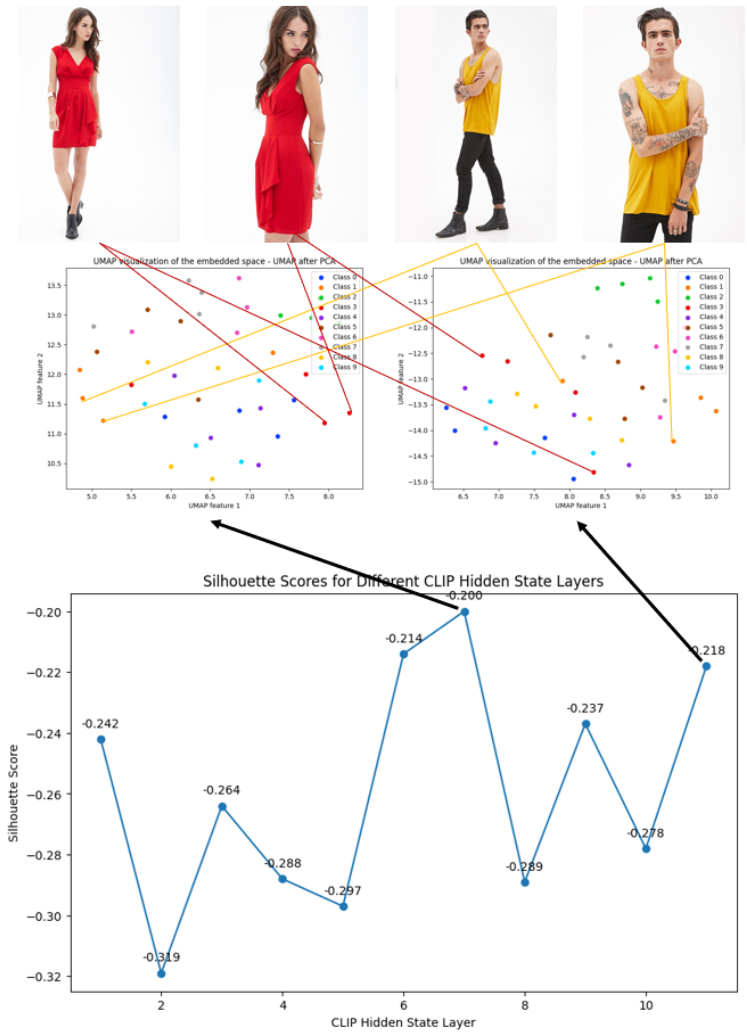

Figure 4.4: Encoding distribution maps for each layer of the CLIP Vision Encoder are presented. A small validation dataset was used, containing 10 sets of different types of clothing data, with each set comprising four images of varying angles and poses. These images were encoded through the CLIP Vision Encoder, and the resulting encodings were dimensionally reduced via PCA and then visualized in a 2D UMAP space. (The arrow indicates the 2D position of images embedding the UMAP space.) We observed that the silhouette scores for the fifth layer were relatively higher, leading to the inference that the fifth layer possesses the most robust capability to distinguish features of clothing. (The arrow indicates the embedding distribution of the specified layer.)

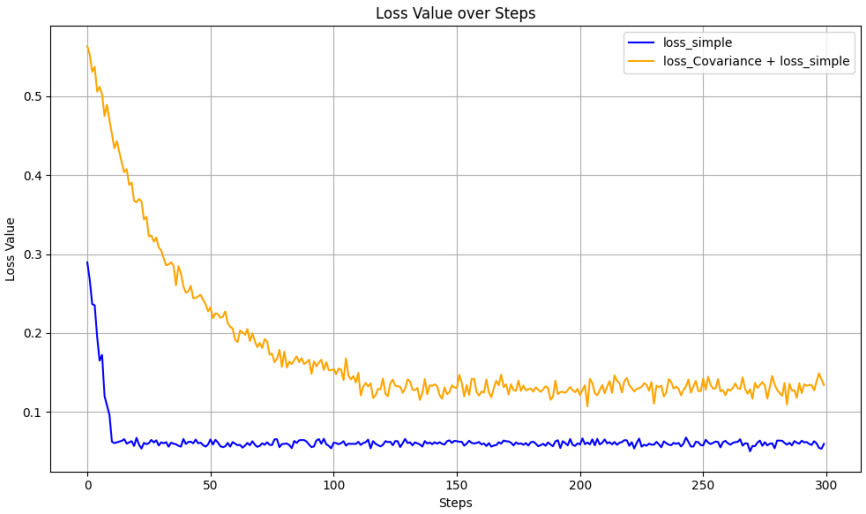

Figure 4.5: The graph compares the convergence trajectories of a simple training loss L_{Simple} against the combined $L_{Covariance+simple}$ approach. Despite the initially higher aggregate loss values of the latter, it demonstrates sustained effective convergence, indicating a more robust learning process over iterations.

of the model. Therefore, to address this limitation and enhance the fine-tuning process, we introduce an additional Covariance-based Loss, $L_{Covariance}$, designed to disentangle the information between pose and style features and promote a more nuanced feature representation.

$$L_{Covariance} = \frac{1}{N} \sum_{i=1}^{N} \left((Z_{pose}^{(i)} - \overline{Z_{pose}}) \cdot (K^{(i)} - \overline{K}) \right) \tag{4.5}$$

where

- N is the number of samples.
- $Z_{pose}^{(i)}$ is the i-th sample of the pose feature.
- $K^{(i)}$ is the i-th sample of the combined style features $K_c + K_{global} + K_{local}$.
- $\overline{Z_{pose}}$ and $\overline{K}$ are the mean values of Z_{pose} and K respectively.

4.3.4 Cross-Pose Style Integration Training (CPSIT)

Consider a dataset $\mathcal{D}$ composed of style-centric image subsets, where each subset contains stylistic variations across multiple poses. For instance, subset $\mathcal{S}_1$ comprises images $\{s_{11}, s_{12}, s_{13}, ..., s_{1n}\}$, each illustrating a unique pose of the same subject or artistic style. Within the CPSIT framework, we may extract

the style representation from s_{11} and merge it with the pose information gleaned from s_{1n}. The model is trained to generate an image s'_{1n} that exhibits the pose characteristics of s_{1n} while infused with the style elements of s_{11}. This process is iteratively performed with various unpaired combinations within the subset, enforcing a versatile internal representation of style and pose that transcends the rigid associations found in conventional paired training datasets.

By disentangling and recombining these features through CPSIT, the generative model becomes adept at synthesizing images that are not restricted by the availability of pose-style matched pairs, thereby granting it a formidable capacity to handle a myriad of real-world applications where such paired data is scarce or nonexistent.

4.4 Experiment

4.4.1 Experimental Setup

Our empirical studies were conducted using the Stable Diffusion (SD) framework, version 1.5. For image encoding, the OpenCLIP ViT-H/14 [8] was employed. The SD architecture comprises 16 cross-attention layers, to each of which we introduced an additional image cross-attention layer, enhancing its multimodal processing capabilities. For the dataset, we utilized the paired image and text description data from the DeepFashion MM [96, 98] dataset. For the pose, we employed OpenPose [64] to detect the skeleton pose data for all the images, removing joints not captured in non-full-body photos. In total, there are approximately 24,000 data pairs, which we divided into a training set and a test set at a ratio of 80% and 20%, respectively. The total count of trainable parameters introduced by our HSI module, inclusive of the resampler network and all the adaptation modules, stands at approximately 75 million. This augmentation renders the HSI not only effective but also computationally efficient. The integration and implementation were facilitated by the HuggingFace diffusers library [99], and for expedited training.

4.4.2 Training Configuration for HSI

Our HSI model was trained exclusively on a single NVIDIA A6000 GPU. Optimization was performed using the AdamW optimizer [100] with a learning rate set to 1×10^{-4} and a weight decay parameter of 0.01. For input preprocessing, images were resized such that their shortest dimension was 512 pixels, followed by a center crop to obtain a uniform resolution of 512 $\times$ 512. In order to implement classifier-free guidance, dropout strategies were applied during training:

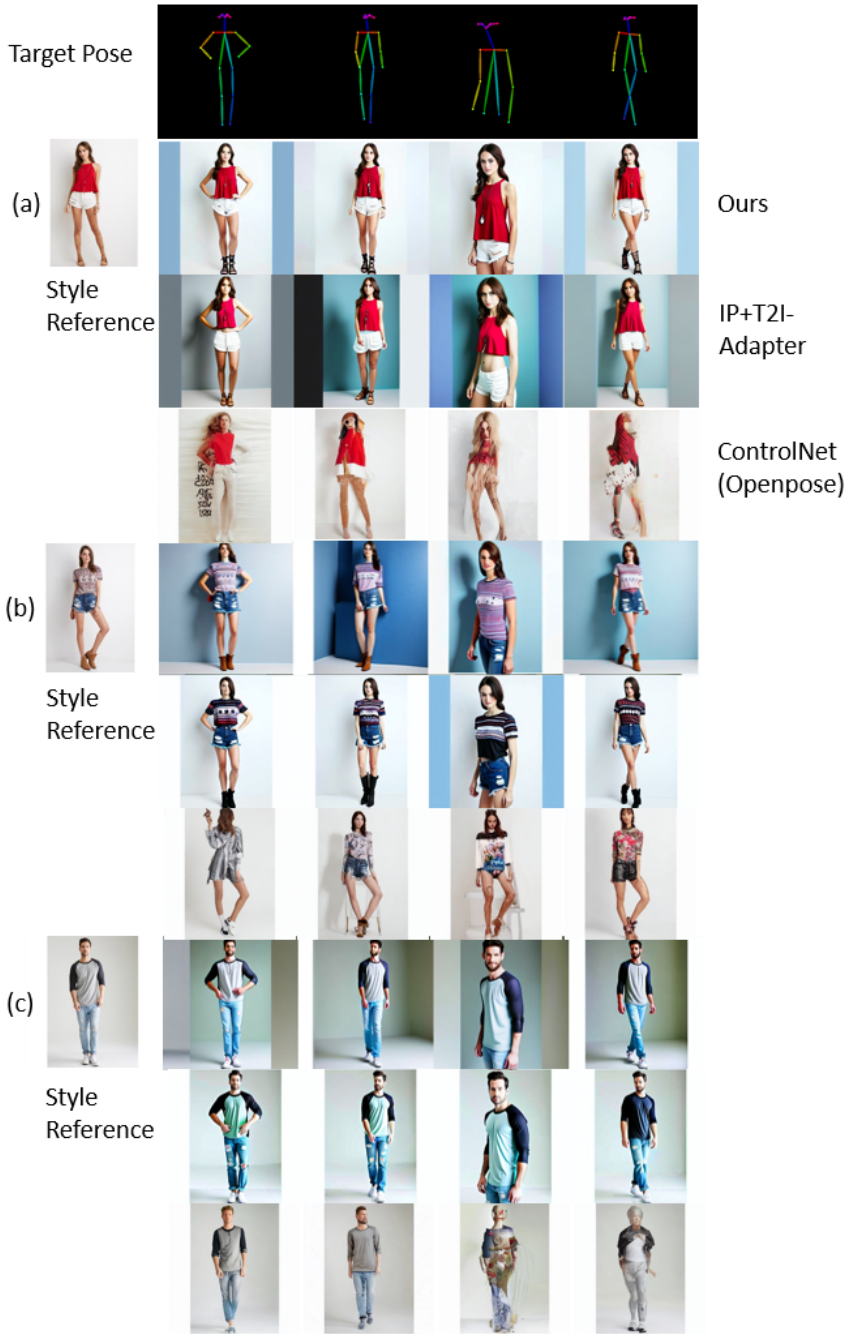

Figure 4.6: The comparative results of pose&style image generation from purposed method, T2I-adapter+IP-adapter, and ControlNet.

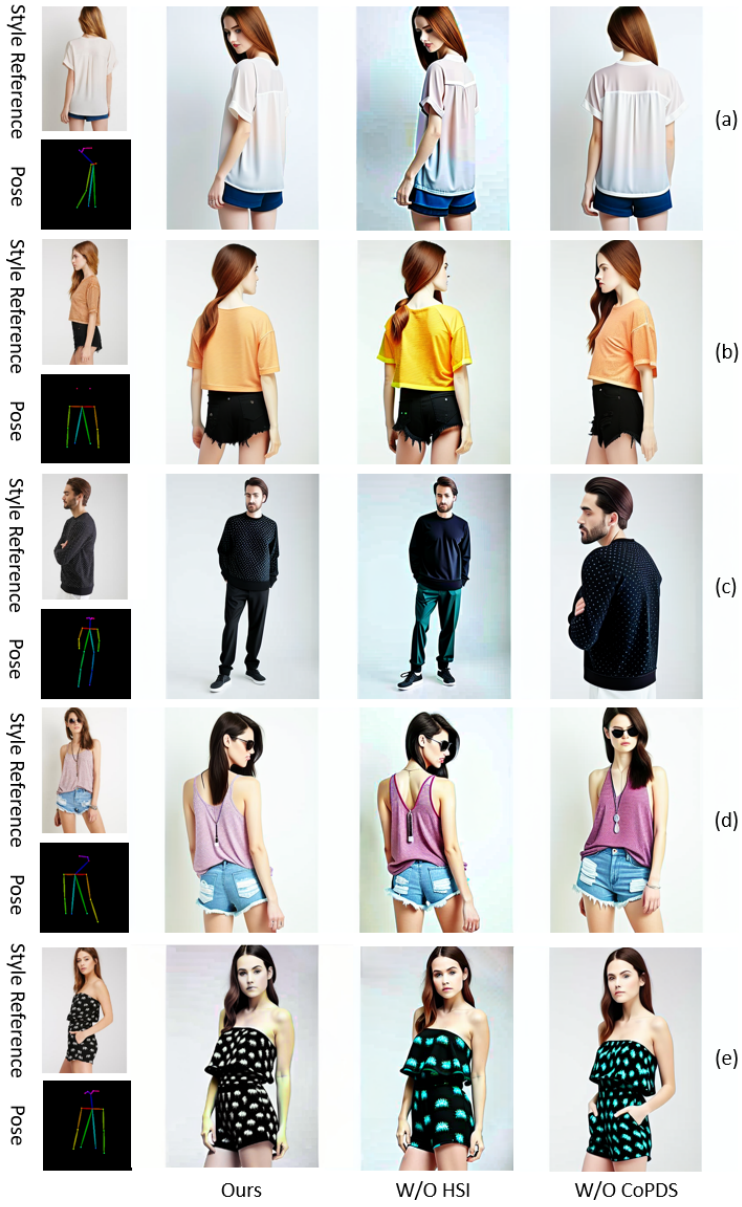

Figure 4.7: The results of ablation study. From left to right, the sequence is as follows: results of the proposed method, proposed method with Hierarchy Style Injection (HSI), and proposed method without Covariance-based Pose and Style Decoupling (CoPSD).

individual text and image inputs were dropped with a probability of 0.05, and simultaneously dropping of both text and image also occurred with the same

probability. For the inference process, a Denoising Diffusion Implicit Models (DDIM) sampler with 50 steps was employed, and the guidance scale was set to 7.5. When generating images based solely on image prompts, text prompts were set to null, and the interpolation weight λ was fixed at 1.0.

4.4.3 Comparion Study

To verify the effectiveness of the purposed method, we compared the generative performance of our results with those of other methods that also fine-tune based on diffusion models, specifically ControlNet [26] and IP+T2I-Adapter [27,40], on the task of pose & style to image synthesis. To ensure a fair comparison, we standardized all models to finetune the pretrained Stable Diffusion v1.5 model. We utilized the DDIM sampling method with 50 steps and employed the classifier-free guidance technique with a scale set to 7.5. The sampling comparison was conducted under the condition of having bidirectional text prompts empty. As demonstrated in Figure 4.6, our method yields consistently stable results in controlling both pose and style. The IP+T2I-Adapter shows instability in some detailed controls, such as the significant alteration in the color of patterns on clothing, as seen in Figure 4.6 (b, c). On the other hand, ControlNet, lacking inherent capabilities for pose and style decoupling, produces inferior results when pose and style references come from different sources. The generated style and pose closely resemble the input style reference, often accompanied by numerous artifacts.

As shown in Table 4.1, we have also calculated CLIP-T, CLIP-I, and LPIPS metrics for the performance of the aforementioned models on the DeepFashion dataset. CLIP-T refers to the textual component of CLIP, which processes and understands textual descriptions, and CLIP-I refers to the image component of CLIP, which processes and understands visual inputs. LPIPS, on the other hand, indicates the degree of difference between the input image and the target image. Based on these scores, we find that the results of the Ours W/O CoPSD group are best in terms of the scores for CLIP-T and CLIP-I. We speculate that any fine-tuning action affects the hidden state of the pretrained Diffusion Model, invariably resulting in some loss of the original model's text-image relationship, whereas HSI ensures a greater degree of the original text-image consistency due to a better understanding of the semantic information of image details. On the other hand, the LPIPS scores show that for better pose condition control, T2I, Ours W/O HSI, and Ours perform better. Based on the observations above, we believe that our results are superior to those of IP-T2I-adapter and ControlNet, particularly in terms of style control, especially the control of stylistic details, as well as pose control.

Figure 4.8: The adaptability of HSI in different customized diffusion models, such as Realistic Vision 2.0 [4] and Anything 4.0. [5]

Table 4.1: Quantitative comparison of the proposed HSI method on DeepFashion Test set. The best scores are shown in **bold**.

	Nums of Parameters	CLIP-T↑	CLIP-I↑	LPIPS↓
IP-adapter (Plus)	29M	0.559	0.795	0.321
T2I-adapter (Openpose)	39M	0.412	0.592	0.265
ControlNet (Openpose)	365M	0,501	0.676	0.319
Ours W/O HSI	51M	0.547	0.758	0.247
Ours W/O CoPSD	57M	**0.562**	**0.797**	0.302
Ours	81M	0.513	0.778	**0.208**

4.4.4 Ablation Study

In the ablation study, we tested the results with and without HSI, as well as with and without CoPSD. As illustrated in Figure 4.7, the generation results that utilized the Hierarchy Style Injection (HSI) technique preserved more details from the style reference inputs, particularly the density and color of the clothing patterns, as exemplified in Figure 4.7 (e). Conversely, the outputs without Covariance-based Pose and Style Decoupling (CoPSD) displayed unstable pose control, as evidenced by the ineffective management of the pose. We infer this instability is due to the fact that during the Cross-Pose Style Integration Training (CPSIT), while the model reinforced style consistency, the input style references did not match the target pose, consequently disrupting the expression of the pre-trained T2I-adapter (openpose-condition) residual hidden state upon fitting. The results of these ablation studies confirm the effectiveness of both HSI and CoPSD. As we can observe, the proposed HSI performed well in controlling the detailed pattern of the clothing style, and the CoPSD assisted in disentangling the pose & style variance to ensure the consistency of both pose & style reference.

4.4.5 Performance on Different Pretrained Models

The proposed HSI module has approximately 78M parameters and can be adapted to all personalized pre-trained SD models. As illustrated in Figure 4.8, we showcase the generative results of HSI on SD1.5 [101], Realistic Vision 2.0 [4], and Anything 4.0 [5]. It is evident that even across different personalized diffusion models, HSI can stably control the input style and pose while performing style transfer corresponding to the specific model.

Figure 4.9: The failure cases of color conditioning (a). The artifact of face reconstruction (b).

4.5 Limitation & Future Work

As shown in Figure 4.9 (a), our model shares a common issue with other image-to-image diffusion models, which is a lack of sensitivity to color control. This is due to our reference inputs being encoded through the CLIP Vision Encoder and aligned with semantic information embeddings before being used as conditional inputs in the denoising U-Net. We speculate that the language model's capacity to express color is inherently vague. For example, the word 'magenta' could visually lean more towards purple or red, and 'light blue' could mean higher brightness or lower saturation in visual terms. Consequently, visual color information does not map well onto the visual cognitive space. In future work, we plan to explore this issue further and propose solutions, such as training an additional color encoder specifically for controlling the denoising of color.

In addition, as shown in Figure 4.9 (b), since our proposed method does not fine-tune the SD-based first-stage VAE (in this case, a VQVAE [102–104] as employed by the paper), similar to other SD-based models, generation of faces, especially when only an image reference is provided without a text prompt, can result in artifacts. This issue typically requires additional negative prompt guidance or a face refinement module to post-process and correct the output. However, we believe that in our future work, this problem could be better addressed by incorporating an additional HSI module dedicated to face refinement.

4.6 Conclusion

In this work, we introduce a fine-tuning strategy for a large-scale pre-trained text-to-image diffusion model, which resulted in a pose & style-to-human generative model. We proposed the Hierarchy Style Injection (HSI) method after analyzing the performance of each cross-attention layer in the denoising U-Net and the feature distribution of the intermediate hidden states of the CLIP Vision encoder, enabling the model to output samples that are consistent with the input style both globally and locally. The Covariance-based Pose and Style Decoupling (CoPSD) technique was then applied to decouple the input style and pose, allowing the model to control various modal input features more stably. Finally, our experiments confirmed the superior performance of our model compared to similar approaches, and due to the modular nature of HSI, it can be applied to a variety of personalized pre-trained diffusion models.

In conclusion, the HSI human image generation method, due to its plug-and-play capability with various pretrained SD models, exhibits relatively high diversity in generated results (Variability). However, since HSI requires users to specifically provide style references for clothing and poses, its tolerance for vague inputs (Flexibility) is somewhat reduced compared to *DualMotion* discussed in Chapter 3. Furthermore, HSI does not demand high precision in editing details like facial features and clothing folds, so its precision (Precision) remains relatively low.

Chapter 5

High-Fidelity Face Image Synthesis with Sketch-Guided Latent Diffusion Model

In the previous chapter, we introduced a method for generating human images by fine-tuning a pre-trained stable diffusion model with inputs of style and pose. We noted that the pre-trained VAE, especially when reconstructing faces, is prone to producing artifacts, and yet facial details are particularly crucial for the overall quality of the human image.

In this chapter, we introduce a diffusion model-based system designed for the synthesis and modification of facial images. The subtle perception humans possess regarding the intricacies of others' facial details is of particular interest. Subtle facial expressions can unconsciously convey profound sentiments, and slight variations in facial features, such as the size of one's eyes, can yield the perception of entirely distinct individuals. Consequently, when users employ generative models to create or modify facial representations, it is paramount that the model closely adheres to their intent—factors like hairstyle, facial structure, and the precise positioning of facial features are indispensable. Furthermore, during the modification process, while the model endeavors to alter specific segments based on user input, it is imperative that the unaltered regions remain consistent and unaffected. To address these nuances and meet the outlined requirements, instead of fine-tuning the cross-attention layers of a pretrained denoising U-Net in Chapter 4, we retrained a denoising U-Net from scratch. In this chapter, we introduce a novel method: High-Fidelity Face Image Synthesis with a Sketch-Guided Latent Diffusion Model.

5.1 Introducion

Generating images from simple monochrome sketches, particularly of human faces, stands as a cornerstone challenge in the domain of image-to-image translation. Such a capability is pivotal for an array of applications, including character

design in digital media and tracking individuals in security contexts. Yet, the inherently sparse nature of single-channel sketch data complicates the extraction of robust and generalized features. Moreover, amassing a well-matched dataset, which pairs artisan sketches with their corresponding real-world photographs, proves to be an arduous and time-intensive endeavor.

Compounding these challenges, synthesis models grapple with interpreting monochrome sketches that often bear superfluous semantic content. Such sketches frequently encapsulate discrete facial components and attributes, ranging from distinct facial features to expressive nuances, and even additional embellishments such as accessories and hairstyles.

GAN-based generative models [105, 106] are one of the feasible solutions for sketch-to-image generation based on semantic mask annotated datasets [107, 108]. Although they allow users to arrange facial semantics (i.e., regional-only conditions), many details may be lost or arbitrarily synthesized, such as wrinkles and mustaches. Instead of applying semantic masks, the other previous GAN-based models [7] trained using sketch-face paired datasets can directly generate (and edit) face images from monochrome sketches. However, they are unsuitable for handling local geometrical details such as accessories and expressions since no semantic information was directly specified in rough monochrome sketches. More recently, the diffusion model (DM) [15, 109, 110] and Contrastive Language-Image Pre-training (CLIP) [111] have achieved tremendous success on the text-to-image task. However, in the case of image-to-image, especially sketch-to-image, their system requires not only image input but also appropriate text inputs, and may not generate desired images, as shown in Figure 5.1. The other conditioning-guided DM-based models such as ILVR [37] and SDEdit [38] approached the image-to-image task by inputting an RGB image reference to control the synthesis. However, it is generally difficult to specify image details after noise injection and resampling of the query input.

To engender generative models with the aptitude to extract more precise information from paired sketches, we introduce the Sketch-Guided Latent Diffusion Model (SGLDM). This architecture, rooted in the principles of Latent Diffusion Models (LDMs), is meticulously trained on a dedicated sketch-face dataset. Owing to the LDM's remarkable prowess in rendering flexible and high-fidelity inferences under a gamut of conditions, it naturally forms the backbone of our sketch-guided image synthesis paradigm.

A pivotal facet of our approach lies in the deployment of a Multi-Auto Encoder (AE). This is strategically designed to transmute query sketches from the granular pixel space to a more abstract feature map, situated in the latent domain of image features. Such a transformation not only mitigates the dimensionality of the sketch input but also assiduously retains the geometric nuances associated with intricate facial details.

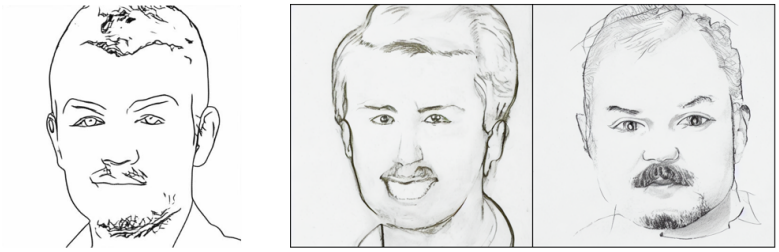

Figure 5.1: An example of implementing a LDM-based model, *stable diffusion* [6] with pre-trained weights. Although we inputted a single sketch (left) and texts (e.g., '*a face photo*' or '*a portrait*'), the generated results are not colored images but monochrome sketches, and do not reproduce the contours of the input sketch.

Comprehending the multifaceted intricacies of sketch and image domains, we adopt a bifurcated training process to ascertain optimal distribution mappings between the two. A salient observation is that individual predilections towards specific facial regions invariably introduce a spectrum of abstraction levels in the input sketch. For instance, while some individuals may exhibit a proclivity for eye details, others might gravitate towards the mouth.

Given such variations, it becomes imperative to account for this heterogeneous abstraction in sketches. To this end, we introduce *Stochastic Region Abstraction* (SRA) – an innovative data-augmentation methodology aimed at fortifying the resilience of SGLDM. It's noteworthy that our sketch data is meticulously extracted from the renowned Celeba-HQ repository, leveraging state-of-the-art (SOTA) sketch simplification techniques [112, 113].

Empirical evaluations validate the advance of our approach, with SGLDM demonstrating the capability to manifest naturalistic face images, accommodating a diverse range of sketch details. Furthermore, our model affords users the latitude to conjure desired face renditions, spanning a resolution of 256×256 pixels, replete with a myriad of expressions, facial accessories, and hairstyles, all guided by a monochrome sketch (elucidated in Figure 5.2).

Distilling our research, we underscore our primary contributions as elucidated below:

- We proposed SGLDM, a sketch-input-only model combining the Multi-AE and DM, and trained a denoising U-Net from scratch.
- SGLDM is trained via a two-stage training process to synthesize faces with high quality and input consistency.
- We introduced SRA, a data augmentation strategy for synthesizing convincible faces from input sketches at different levels of abstraction.

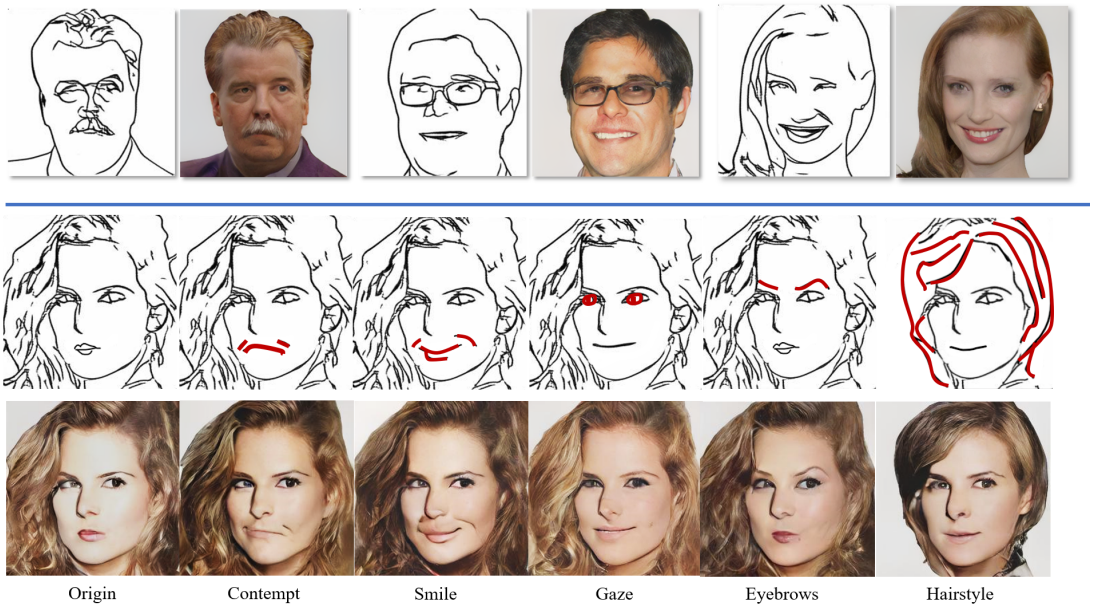

Figure 5.2: A Sketch-Guided Lantent Diffusion Model (SGLDM) synthesizes high-quality face images with high consistency of input sketches. SGLDM enables users to simply edit face images such as different expressions, facial components, hairstyles, etc. The edited strokes are highlighted in red.

- We verified the SGLDM achieves superior scores in various metrics compared to SOTA methods and it is sufficiently robust to generate the intended face images.

5.2 Sketch-based Image Synthesis

5.2.1 GAN-based

The field of image synthesis from sketches has witnessed significant exploration over the past decade. From its inception, the problem of sketch-to-image translation has been tackled as an image-to-image transformation task. Researchers have endeavored to train deep learning-based networks to bridge the gap between monochromatic sketches and full-color RGB images. Several supervised Generative Adversarial Network (GAN)-based generative models, such as Pix2Pix [114] and Pix2pixHD [115], rely on paired sketch-image datasets, which are created by extracting the edge information from real images to facilitate model training.

To enhance the efficacy of translating the image domain into the sketch domain,

the construction of a substantial corpus of paired sketches and photographs is imperative. Consequently, datasets like Sketchycoco [116], which categorize objects into distinct classes, have been introduced. However, when it comes to facial sketch image datasets, the availability is notably limited, as exemplified by datasets such as CUHK Face Sketches [117, 118].

On the other hand, unsupervised image-to-image translation methods, such as CycleGAN [114] and DualGAN [119], have been explored in other works. More recently, with the burgeoning advancement in disentangled representation within StyleGAN's $w+$ space, sketch-to-sketch translation has been treated as a style transfer task, illustrated by methods like DualStyleGAN [120] and Pixel2Style2Pixel [94]. Furthermore, akin to the recently popular text-to-image generative models, GAN-based approaches also permit users to generate and edit images via textual and sketch inputs, such as [121, 122]

Nonetheless, end-to-end GAN-based models have been associated with issues like unstable training and susceptibility to overfitting on specific datasets. These challenges restrict the diversity and quality of synthesized results. Hence, drawing inspiration from the notable performance of Latent Differential Models (LDM) in conditional image synthesis tasks, we propose SGLDM as a solution for achieving high-quality face synthesis with enhanced input consistency.

5.2.2 DM-based

In recent times, diffusion and score-based models have emerged as formidable contenders in the realm of image synthesis. A fundamental component of these models is the U-Net architecture [123], which has been lauded for its excellence in fostering diversity, ensuring quality, stabilizing training, and offering module extensibility.

Noteworthy advancements have been presented in previous studies, such as [15, 109], which have demonstrated superior performance, particularly in the domain of unconditional image synthesis. However, a significant impediment remains the hefty computational costs, which subsequently constrain the resolution of the images produced. In a serendipitous turn of events, contributions by [6] have provided a solution. In their approach, images are initially encoded from a high-dimensional RGB space to a more manageable, low-dimensional latent feature space.

Subsequently, this latent representation is employed to navigate both the forward and backward diffusion processes. Additionally, the architecture's commendable modular extensibility equips it to handle a plethora of image-to-image tasks. This includes but is not limited to, image inpainting, semantic mask-to-image translation, and layout-to-image generation, as elucidated by [6].

While certain methodologies have focused on altering the network architecture of diffusion model (DM)-based designs, alternative approaches start from ILVR [37] and SDEdit [38]. Up to more recently represented by ControlNet [26] and T2I-Adatper [27] have chosen a different path. These strategies involve fine-tuning the models with supplementary plug-in condition modules. Alternatively, during the sampling process, they incorporate extra constraint loss functions to govern the sampling procedure. This concerted effort has yielded impressive results, enabling models to excel in the task of generating high-quality images from sketches. Intriguingly, their method necessitates a blurry RGB reference, which serves the dual purpose of iteratively guiding the sampling process and acting as an input reference. Despite their efforts, the delineation of intricate image details was compromised, primarily attributable to the conditioning's inherent blurriness. Meanwhile, when considering the sketch-to-image task, a paramount challenge emerges: the monochromatic sketches inherently possess a dearth of semantic information. Consequently, executing the sketch-to-image transformation via Diffusion Models (DM) invariably demands supplementary inputs, such as auxiliary text prompts, to compensate for this information void.

5.3 Method

5.3.1 Overview

In our endeavor, we aim to craft high-fidelity facial images that faithfully mirror the intrinsic characteristics of the input sketch. We postulate that the feature distribution associated with monochromatic sketches within our dataset exhibits greater irregularities and sparsity when juxtaposed against that of full-colored RGB images. This speculation is underpinned by the observation that a simultaneous training of a sketch embedding, meant to bridge the chasm between the sketch and image domains, can potentially yield a discontinuous feature distribution, as the illustration depicted by the dashed curve in Figure 5.4.

In response to this challenge, we architected a bifurcated training methodology, ensuring a more fluid and optimized mapping across the sketch and image domains. This optimized mapping is visually represented by the solid curve in Figure 5.4. A thorough dissection of this methodology and its underlying nuances is presented in Section 5.4.

5.3.2 Preliminaries

Since the previous chapter has already introduced the preliminaries of DM-based knowledge, here we briefly touch upon and enumerate the equations relevant to

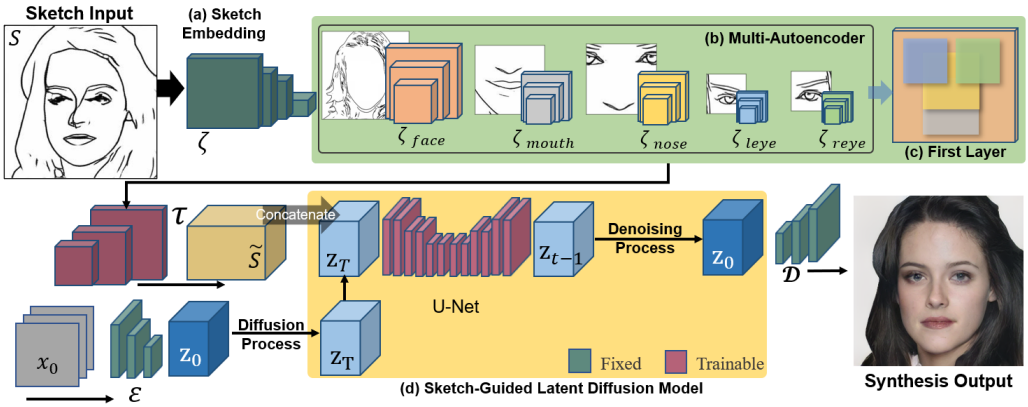

Figure 5.3: The framework of SGLDM. In the sketch embedding stage (a), given a sketch input S, a pretrained sketch encoder ζ encodes S into a feature (c) $= \{\zeta_{leye}, \zeta_{reye}, \zeta_{nose}, \zeta_{mouth}, \zeta_{face}\}$. The decoder τ then decodes S' into a feature map $\tilde{S}$. In the latent denoising stage (b), the random latent code Z_T is concatenated by the feature map $\tilde{S}$ and denoised to Z_0 by a U-Net. Finally, the output latent code Z_0 is decoded by a decoder $\mathcal{D}$ to the final output.

this study. The origin loss of training DM is to calculate the simplified variant of the variational lower bound:

$$L_{DM} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(x_t, c, t)\|^2] \tag{5.1}$$

where x_0 denotes the authentic data, which are augmented with a conditioning element, represented as c. The diffusion process temporal progression is captured by t, which ranges within $[0, T]$. This acts as a chronological gauge, signaling the evolution of the diffusion process across its steps. The noisy data at a particular time step t are symbolized by x_t, a function of the genuine data $x0$, the Gaussian noise ϵ, and the predefined coefficients α_t and σ_t. Specifically, $x_t = \alpha_t x_0 + \sigma_t \epsilon$ articulates the amalgamation of the real data and noise, moderated by α_t and σ_t, the roles of which are pivotal. They are not arbitrary but are systematically defined functions of t, and their values influence the trajectory and intensity of the diffusion process. After successfully training the model ϵ_θ, it is then empowered to generate visual content, or images, starting from arbitrary noise. This generation is not instantaneous, but unfolds iteratively, resembling the incremental character of the diffusion process.

In a more recent development, a method called LDM, as introduced by Rombach and colleagues [6], has emerged with the aim of mitigating computational costs. The rationale behind this approach hinges on the observation that, even after passing through the neural network of the AE model, certain features that

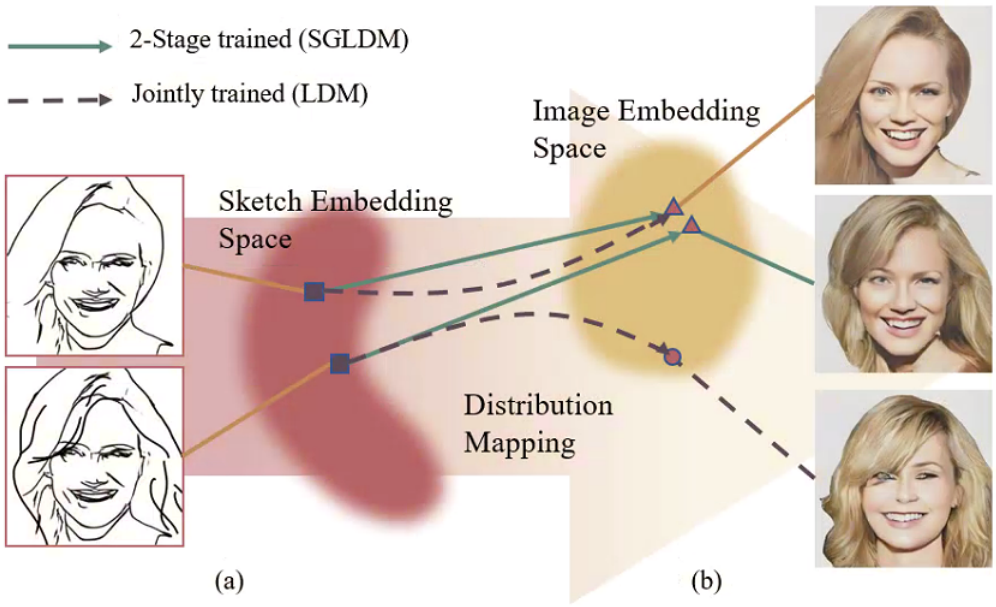

Figure 5.4: The illustration of feature distribution mappings between the jointly-trained conditional embedding (dashed line) and the separately-trained conditional embedding (solid line), from the sketch (a) to the image (b) domain.

contribute to perceptual intricacies and semantic significance remain embedded within the latent code.

The LDM technique involves the utilization of a pre-trained encoder denoted as ε. This encoder is tasked with encoding an image x residing in a high-dimensional RGB space, represented as $x \in \mathbb{R}^{H \times W \times 3}$, into a lower-dimensional latent code z, situated in $z = \varepsilon(x) \in \mathbb{R}^{h \times w \times 3}$. Note that H, W, and h, w are the height and width of the image and the feature map respectively. Subsequently, a pre-trained decoder D is employed to reverse this process, effectively generating images from the latent code, denoted as $\tilde{x} = \mathcal{D}(z)$. The significance of this transformation lies in its potential to facilitate a switch in the loss-term L_{LDM}.

$$L_{LDM} = \mathbb{E}_{x_0, \epsilon \sim N(0,I), c, t}[\|\epsilon_t - \epsilon_\theta(z_t, c, t)\|^2] \tag{5.2}$$

5.4 Sketch-guided Latent Diffusion Model

5.4.1 Framework

In the context of face synthesis, our objective is to generate facial images based on a single provided sketch input. To achieve this, we treat the sketch as a guiding condition for the model during the denoising process. Inspired by the seminal work of Chen et al. in DeepFaceDrawing [7], our approach leverages a region-specific *Multi-AE* architecture for enhanced facial feature extraction. This methodology allows for distinct and precise processing of various facial regions, facilitating a more nuanced and detailed representation in facial feature analysis. The framework of our approach, denoted as SGLDM, is illustrated in Figure 5.3 (a, d). Drawing inspiration from the work of Rombach and colleagues [6], we also incorporate an LDM to optimize computational efficiency.

Incorporating our sketch-condition pairs, the training loss L_{SGLDM} for the conditional LDM can be expressed as follows:

$$L_{SGLDM} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(z_t, t, \tau_\theta(\tilde{S})\|^2] \tag{5.3}$$

In this formulation, $\tilde{S}$ represents a sketch feature that has been encoded by a pre-trained sketch encoder, denoted as $\zeta(S)$, operating on the input sketch S. The τ_θ function serves as a decoder, responsible for estimating a conditional map that allows for the reversal of the diffusion process applied to $\varepsilon(x)$. It's important to note that both τ_θ and ϵ_θ are concurrently trained.

Instead of solely training a sketch encoder to generate a conditional feature map for Z_T to facilitate denoising, we introduce a "Conditioning Module" by pretraining a Multi-AE network architecture. Drawing inspiration from previous works that segmented the global facial structure into local components for individual networks, as seen in DeepFaceDrawing [7], APDrawGAN [124], and Manga-GAN [125], our overarching encoder ζ comprises five distinct partial encoders, denoted as $\zeta = \{\zeta_{leftEye}, \zeta_{rightEye}, \zeta_{nose}, \zeta_{mouth}, \zeta_{face}\}$, as illustrated in Figure 5.3 (b,c).

For face editing, as opposed to face synthesis, we introduce an original facial input. To facilitate this, we train a VQVAE, which employs vector quantization to enhance the quality of image synthesis by learning discrete latent representations, using our sketch dataset to encode the dilated sketch. Both the original face and the sketch input are inversely masked, with the face being masked by the region designated for editing and the sketch input being masked by the remaining area. Subsequently, we concatenate the two encoded features with an additional binary mask map to train the LDM.

5.4.2 2-Stage Training Strategy

In our approach, we employ a bipartite training regimen, systematically constructed to enhance the generative prowess of our model. Initially, during the 'sketch embedding phase', we place our emphasis on the pre-training of the *Conditioning Module*. The crux of this foundational phase is the optimization of the model parameters, driven by the overarching aim to reduce the aggregate Mean Squared Error (MSE) loss, denoted as $L_{Multi-AE}$. This loss is an embodiment of the discrepancies between the encoded-reconstructed images and their true counterparts, and it emanates from each distinct partial encoder present in our model's architecture. The mathematical articulation of this objective is as follows:

$$L_{Multi-AE} = \|\sum_{\zeta_i \in \zeta} \zeta_i(x) - x\|_2 \tag{5.4}$$

Opting to commence with a pre-training phase focused on the *Multi-AE* as opposed to diving straight into an integrated training regime for sketch encoders—further establishing a conditional feature map for the SGLDM—finds its rationale rooted in two cardinal imperatives:

- **Domain Distribution Alignment:** Our intent is to cultivate a model that more adeptly discerns and maps the relationships between the distinct domain data distributions characterizing sketches and faces. By doing so, we aim to yield a seamlessly integrated domain distribution space, as visualized in Figure 5.4.
- **Computational Efficiency:** Adopting a two-stage training strategy provides computational advantages. Specifically, by decoupling the trainable parameters of models across distinct stages, we streamline and enhance the model optimization process.

In the ensuing phase of our training regimen, we further refine our LDM. In our quest to enhance the versatility of SGLDM across a spectrum of sketches, we introduce what we have coined as the *arbitrarily masking conditional training strategy*. Rooted in the foundational principles of the *Masked AutoeEncoder* as expounded upon by He et al. [126], this methodology revolves around the strategic obfuscation of select portions of the input. This not only challenges but also compels the model to embark on an autonomous quest to restore these deliberately concealed segments. Given the pre-trained architecture of our sketch encoder, denoted by ζ, our strategy delves into the random masking of the conditioning feature map, symbolized as S. This intentional masking is pivotal during the training epoch designed for the denoising U-Net, compelling the model to bridge the occluded segments with learned data, thereby fortifying its predictive prowess.

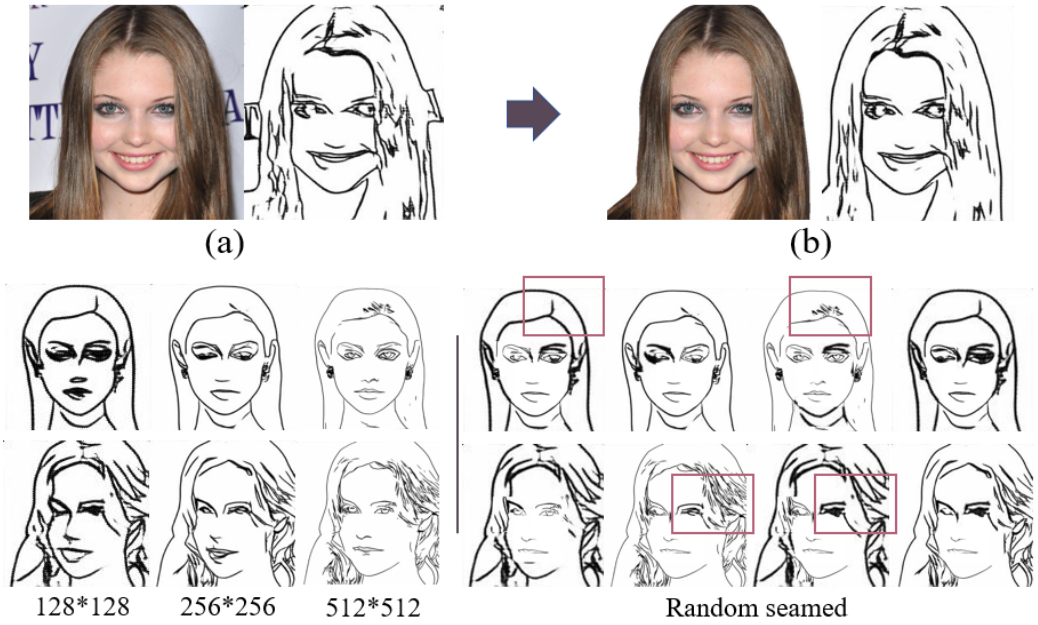

Figure 5.5: The original image in Celeba-HQ and its extracted edge map (a), and the result of paired data after cleaning up the background (b). Sketch simplification results from 3 different resolution faces (left-bottom). And the random seamed data samples (right-bottom).

5.4.3 Stochastic Region Abstraction Data Augmentation

To construct our training dataset, we sourced 10,000 high-definition facial images from the CelebA-HQ dataset [107]. Initially, we processed these images to remove any background, as illustrated in Figure 5.5 (a,b). Subsequently, we employed the *sketch simplification* methodology [112, 113] to produce facial edge maps.

In our pursuit to bolster the capability of the SGLDM system in handling sketch inputs across varying degrees of abstraction, we incorporated the SRA to augment our dataset. We discerned that the abstraction granularity of the generated edge maps was intrinsically influenced by the resolution of the image.

Thus, we undertook a resizing process of the original images to resolutions of 128×128, 256×256, and 512×512, respectively, as exemplified in Figure 5.5 (left-bottom). This step further enriched our sketch dataset. A noticeable variance in abstraction, especially around the regions encompassing the hair and eyes, is spotlighted by the red bounding box. Further enhancing the diversity of our dataset, and in line with our *Multi-AE* approach concerning individual encoder regions, we segmented the edge maps into five distinct sections. These sections were then amalgamated in a randomized manner to produce new edge maps char-

acterized by randomly positioned seams at diverse abstraction levels, as portrayed in Figure 5.5 (right-bottom).

Conclusively, out of the total image set, 8,000 images were earmarked for training, 1,000 for validation, and the remaining 1,000 designated for testing.

5.5 Experiment and Results

We conducted several experiments to verify the quality and sketch input consistency of SGLDM's synthetic face images.

5.5.1 Implementation

To ensure the robustness and efficiency of the SGLDM, our training was conducted on an advanced NVIDIA RTX3090 GPU. The two-phase training can be elaborated as follows:

In the inaugural phase, focusing on *Multi-AE* training, we ran the training process for a span of 500 epochs. For optimization, we deployed the Adam optimizer, with hyperparameters set at $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The selected batch size for this stage was 64. Notably, each autoencoder (AE) shared a consistent latent space dimensionality, which was fixed at 512. Transitioning to the second stage, the core SGLDM was subjected to a training regimen lasting 300 epochs. Again, we favored the Adam optimizer. However, given the nuanced requirements of this phase, we opted for a more fine-grained batch size of 8. Analyzing the feature map architecture, the sketch embedding was designed with 8 channels. Factoring in the additional 3 channels from the LDM latent size, the denoising U-Net was fed an 11-channel latent code. The resulting output from this intricate design culminated in a 3-channel configuration.

5.5.2 Quantitative Comparisons

In our quest to ascertain the efficacy of SGLDM, we benchmarked it against a cohort of leading-edge image-to-image translation algorithms in the domain of sketch-to-face translation. Specifically, we considered the following methods: pip2pixHD [115], pix2pix [114], DeepFaceDrawing [7], pixel2style2pixel (PSP) [94], and Palette [127].

To ensure a level playing field, we retrained the majority of these models on our curated set of 10,000 facial images, extracted from the CelebA-HQ dataset, all under uniform training parameters. Notably, we made a direct application of the pre-existing weights from the DeepFaceDrawing model, specifically optimized for the 512×512 resolution.

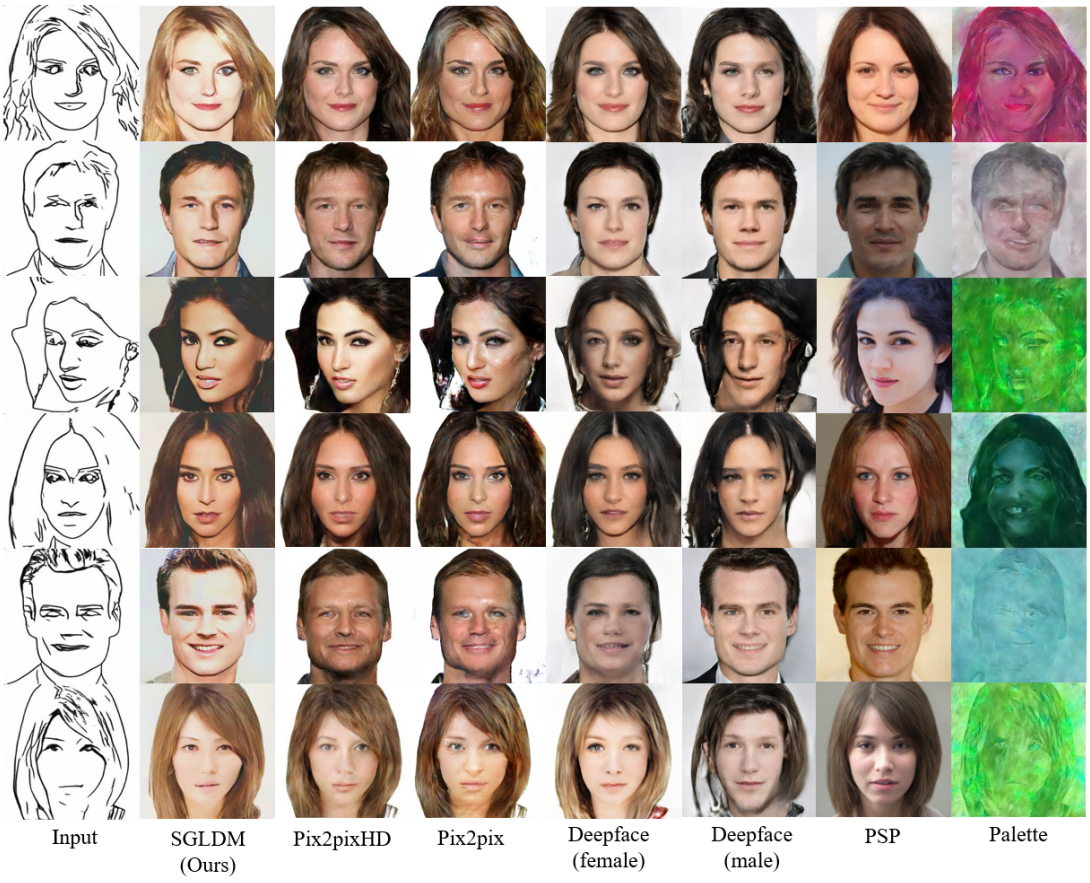

Figure 5.6: Qualitative comparisons of the proposed SGLDM with the SOTAt methods.

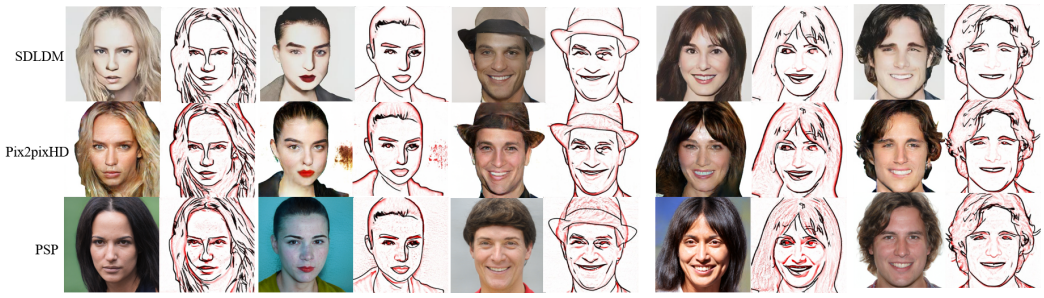

Figure 5.7: Fidelity comparisons of the proposed SGLDM with competing methods.

The aggregate visual output, as evidenced in Figure 5.6, emphatically illustrates that SGLDM adeptly synthesizes facial images that are both strikingly realistic and rigorously faithful to the input sketches. Contrarily, models such as Pix2pix, Pix2pixHD, and DeepFaceDrawing manifested a proclivity to generate images with noise artifacts, particularly evident when the sketches did not depict a front-facing orientation, as can be seen in specific columns. It's worth highlighting that DeepFaceDrawing mandates an additional gender-specific condition for face synthesis, thus, we have provided results for both gender variations.

While the PSP model managed to deliver visually superior results compared to several other contenders, it faltered in maintaining fidelity to the original sketches. On the other hand, Palette, a method rooted in DM-based image translation, could not adeptly generate credible faces when relying solely on sketch input. As of our current understanding, the pinnacle of DM-based methods predominantly leverage inputs beyond monochromatic sketches, often intertwining with techniques such as text-to-image, segmentation-map-to-image, or image inpainting combined with sketch inputs, as seen in works like [128].

Subsequently, we embarked on a comparative analysis featuring SGLDM, Pix2pixHD, and PSP, given their comparable outcomes in terms of fidelity, as elucidated in Figure 5.7.

To provide context, the stark black strokes on the right serve as input sketches, pivotal for the synthesis of the facial images depicted on the left. Superimposed on these black strokes are red strokes, representing the modified renditions of the synthesized images post-application of Adobe Photoshop's *sketch filter* [129].

Upon meticulous inspection of the results, it becomes evident that SGLDM excels in generating immaculate faces, largely devoid of noise while ensuring profound alignment with the initial sketch. However, minor deviations can be observed in intricate facial elements, such as the nasolabial folds. For a richer visual narrative, Figure 5.8 showcases synthesized facial images replete with di-

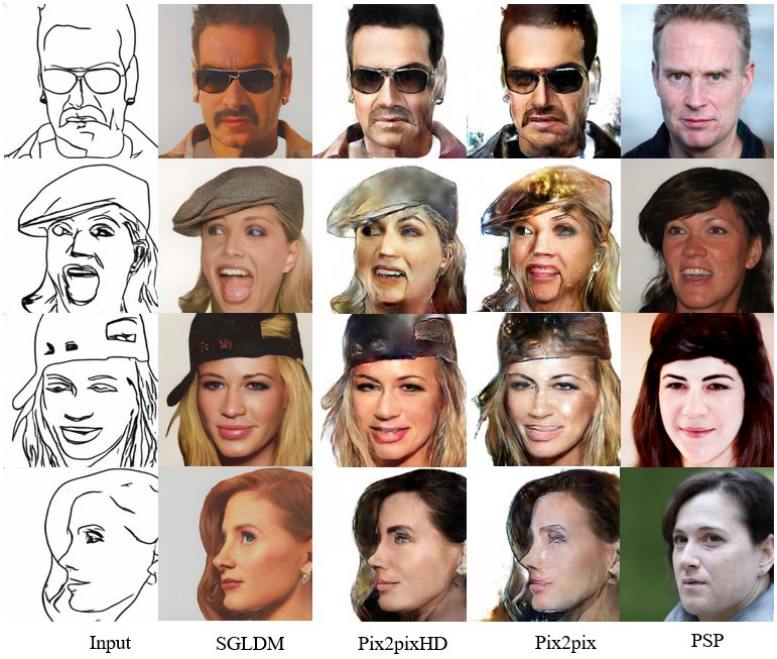

Figure 5.8: Examples of corner cases of sketch input which carried glasses, hat, or side face.

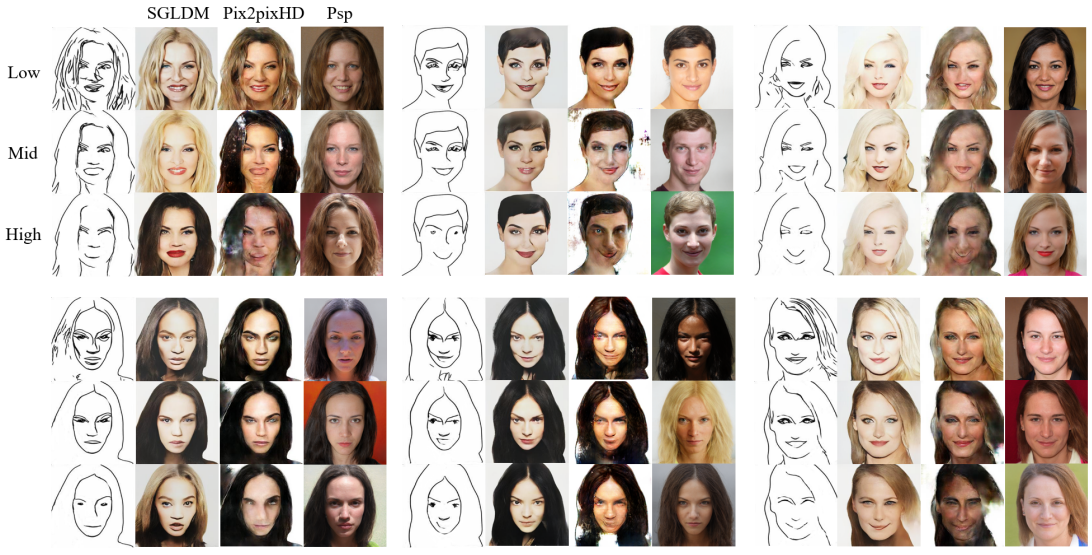

Figure 5.9: The comparison synthetic results in different sketch inputs with three abstraction levels.

Table 5.1: Preference result of user study.

Method	Quality	Fidelity
Pix2pixHD [115]	17%	27%
Psp [94]	37%	2%
Ours (SGLDM)	**46%**	**71%**

verse expressions, ornamental accessories, and varied hairstyles. It's manifest that our method strikes an optimal equilibrium between visual allure and unwavering consistency with the provided inputs.

Furthermore, we conduct a user study to compare the visual aesthetics and input fidelity offered by three methodologies: SGLDM, Pix2pixHD, and PSP. 43 participants, comprised of master's and Ph.D. students with a background in computer science, were asked to participate in the study. It is noteworthy that Pix2pix was omitted from this comparison, given its visual output bears a striking resemblance to that of Pix2pixHD. The cohort of participants was instructed to articulate their preferences, selecting from a triad of synthetic facial images engendered by the aforementioned models, evaluating both the axes of visual splendor and alignment with the input sketch.

Perusing the results delineated in Table 5.1, it becomes evident that facial images manifested by SGLDM command a preeminent preference in terms of congruence with the input. Moreover, in the realm of visual aesthetics, SGLDM's prowess mirrors that of PSP, underscoring its robust capability.

5.5.3 Qualitative Evaluation

In assessing input fidelity, we determined the recall ratio (REC) between the distinctive black and red strokes, as depicted in Figure 5.7. Moreover, given the subtle visual disparities in outputs stemming from varied input sketch resolutions (elaborated upon in Section 5.4.3), we curated input sketches by judiciously erasing certain strokes from the primal sketches, leading to the generation of respective facial images, as showcased in Figure 5.9.

The derived outcomes accentuate SGLDM's robustness, fortifying its capability to adeptly interpret and translate sketches spanning a spectrum of abstraction levels. In our pursuit of comprehensive analysis, we executed an ablation study, juxtaposing metric scores between joint training paradigms and our 2-Stage training methodologies, thereby appraising the efficacy of our SRA strategy. These findings are illustrated in Table 5.2 (lower rows).

Our observations denote that SGLDM, when nurtured through joint training approaches, echoes the performance caliber of Pix2pixHD. Interestingly, while

Table 5.2: **Quantitative comparisons.** We applied the FID (↓) score to measure the synthetic faces quality, the LPIPS (↓) scores to evaluate the consistency between real faces and synthesized results, and a recall ratio (REC↑) to evaluate the input consistency.

Method	Low abstraction			Mid abstraction			High abstraction		
	FID↓	LPIPS↓	REC↑	FID↓	LPIPS↓	REC↑	FID↓	LPIPS↓	REC↑
Pix2pix [114]	53.67	0.20	0.54	59.46	0.23	0.50	63.45	0.28	0.51
Pix2pixHD [115]	51.23	0.18	0.62	53.71	**0.22**	0.55	60.23	0.25	0.53
Psp [94]	83.48	0.29	0.37	83.32	0.26	0.45	85.54	0.28	0.48
SGLDM *joint training*	46.28	0.20	0.65	48.62	0.23	0.54	50.33	0.26	0.51
SGLDM w/o *SRA*	**38.57**	**0.17**	**0.77**	48.87	0.26	0.51	57.76	0.29	0.48
Ours (SGLDM)	43.58	0.22	0.71	**45.46**	0.24	**0.59**	**46.83**	**0.24**	**0.57**

SGLDM, trained sans the SRA and solely on datasets characterized by a singular abstraction level, excelled in translating intricate sketches (indicative of low abstraction), its proficiency plummeted when presented with more abstracted inputs, as evident in Figure 5.10 (extreme right column, descending order).

Conclusively, the insights garnered underscore that bifurcated training combined with astute data augmentation strategies bolster SGLDM's overarching prowess across a diverse array of sketch inputs, as vividly captured in Figure 5.10(penultimate column).

5.5.4 Editing Capability

In our analytical framework, we deeply probed the utility of face editing. Figure 5.11 provides a vivid exposition of selective edits: (a,b) delineate alterations to hairstyles spanning both genders, (c) highlight modifications to earrings, and (d) capture nuanced shifts in facial expressions.

Further, we embarked on a comparative trajectory between facial renditions birthed by the 2-Stage training regimen and those cultivated through a joint training paradigm, as visualized in Figure 5.10. This comparative analysis underscores the heightened resilience of the 2-stage trained SGLDM is more robust than the jointly trained SGLDM, forming a different identity easily after editing, (see Figure 5.10 (third column, upward)). As a result, the SGLDM is sufficiently robust enough to edit the intended face at will using the synthetic results.

In summary, SGLDM distinguishes itself as a strong framework, providing users with the consistency to smoothly shape and fine-tune facial features within its generated outputs.

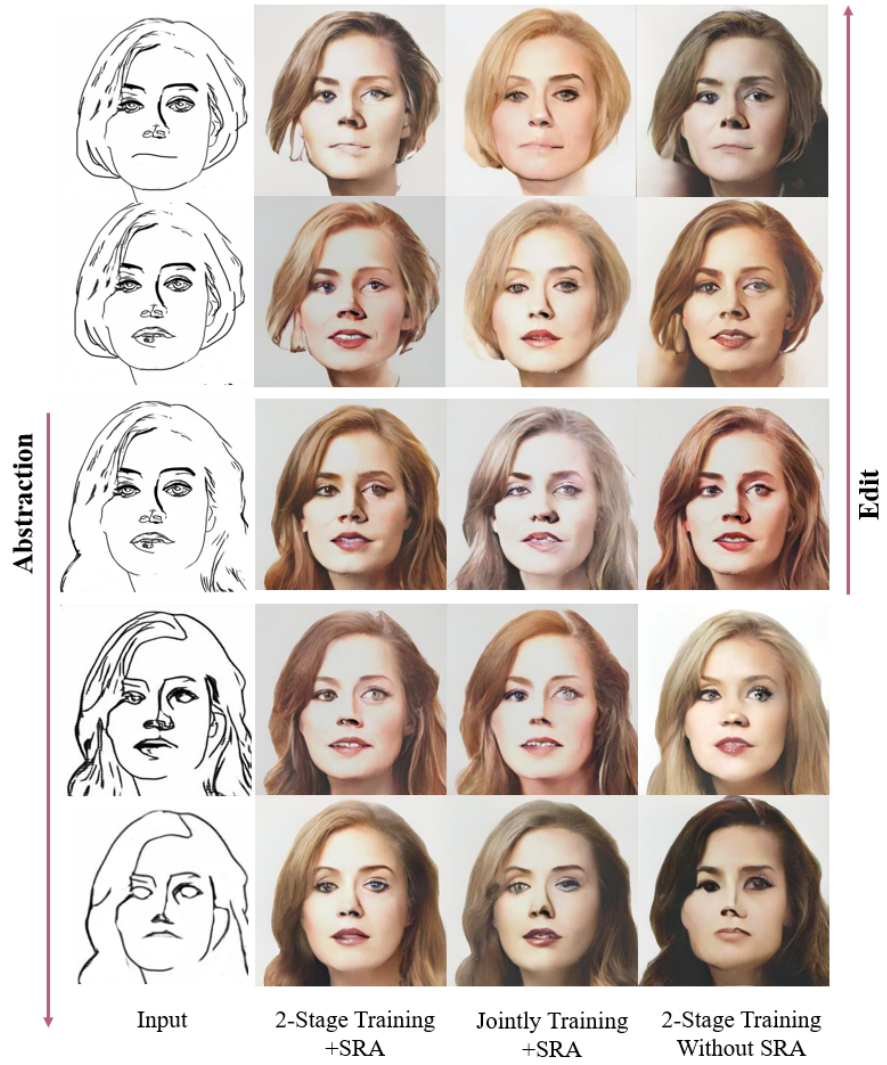

Figure 5.10: The synthetic faces of ablation study.

5.6 Limitations & Future work

While SGLDM commendably retains a high fidelity to input sketches, the resultant synthesis occasionally exhibits an over-reliance on the sketch input. Specifically, the introduction of notably flawed sketches can inadvertently birth noise and anomalies, as exemplified in Figure 5.12. Compared to other SOTA methods, particularly the results from Psp, our approach maximizes the realism of the output samples. However, there is a significant divergence between the output results and the input sketches. Conversely, our results sacrifice the credibility of the images

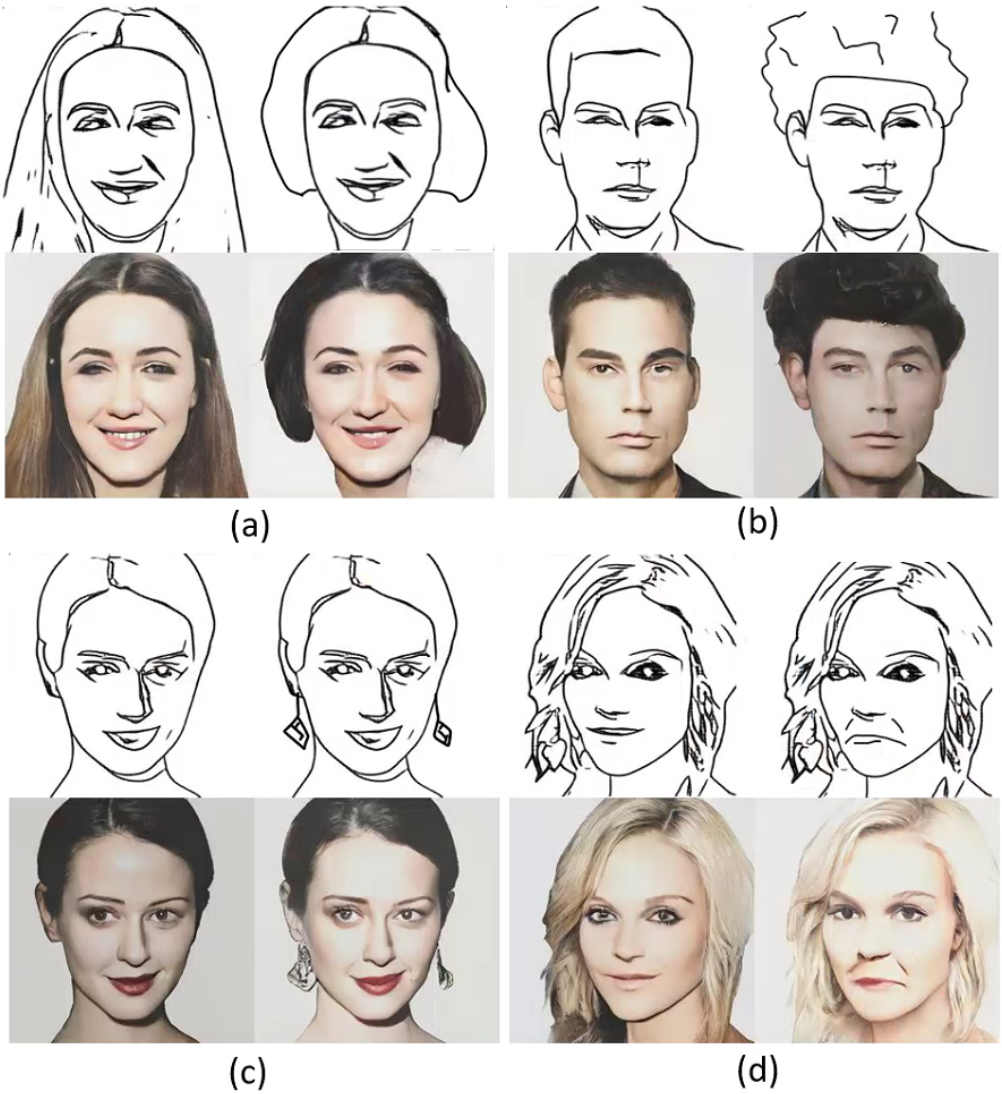

Figure 5.11: Examples of face editing with SGLDM. (a,b) hairstyles, (c) earrings, and (d) expression.

to the greatest extent but ensure the highest level of consistency with the input sketches. To navigate this challenge, future work might contemplate integrating techniques that strike an optimal equilibrium between preserving input consistency and ensuring output realism.

Inspired by the pre-processing technique proposed in DeepPS [130] for enhancing input sketches, we incorporated this approach into our training data. Specifically, we fine-tuned the already trained DiffFaceSketch model with sketches enhanced using this method. As shown in the lower part of Figure 5.13, this

approach compromises to some extent the consistency between input and output to ensure a reduction in artifacts in the output. Although the results still show some discrepancy compared to real samples, the consistency is notably better than other SOTA methods.

Beyond purely monochrome sketches, we envision broadening our methodology to incorporate sporadic color cues, which could potentiate more intricate handling of chromatic nuances in regions such as skin and hair.

Furthermore, our assessment of SGLDM, while rooted in facial synthesis, reveals a framework that potentially lends itself to other sketch-image domains. By simply modulating the training dataset to others like LSUN [131] or AFHQ [132], similar architectures could be devised. Our SRA technique, being versatile, can seamlessly augment these datasets, fortifying the resilience of the respective models.

In this work, although we implemented an LDM-based method to reduce the computation costs, the SGLDM (i.e., the training and the sampling stage) is still computationally heavier than GAN-based models. In the training stage of a 256×256 model, the maximum batch size on a single NVIDIA RTX3090 is 8, while a 512×512 model's maximum batch size is only 1, and in the sampling stage, the average time cost of one image is around 15.2 seconds. Although it can be cut down to 50 sampling steps and takes around 5 to 6 seconds when using the denoising diffusion implicit models (DDIM) sampling strategy, the current implementation is still difficult to incorporate into a real-time interactive graphical user interface (GUI). Furthermore, the latest methods like the Latent Consistency Models (LCMs) [133] significantly reduce the sampling generation time by requiring only 2 to 4 steps for sampling. We plan to reference their approach for application in our SGLDM to enable more real-time interaction in the future.

5.7 Conclusion

In this study, we introduce SGLDM, a new LDM-based framework designed specifically for facial synthesis. This model utilizes a multi-AE mechanism to encode input sketches into feature maps while preserving the complex geometric details of facial features. A key component of our method is the SRA, a novel data augmentation technique that enables the model to handle sketches with varying levels of abstraction.

Our empirical evaluations confirm SGLDM's ability to produce high-quality facial images with a high degree of fidelity to the input sketches. Moreover, our model exhibits a remarkable capacity for editing the generated images, accommodating a wide range of expressions, facial accessories, and hairstyles.

In summary, DiffFaceSketch, due to its high consistency in allowing users to input sketches and output detailed facial features, demonstrates a high level of precision (Precision). However, this high precision results in DiffFaceSketch having a lower tolerance for vague inputs (Flexibility) and reduced diversity (Variability) in generated samples, especially when compared to *DualMotion* and *HSI* discussed in Chapters 3 and 4.

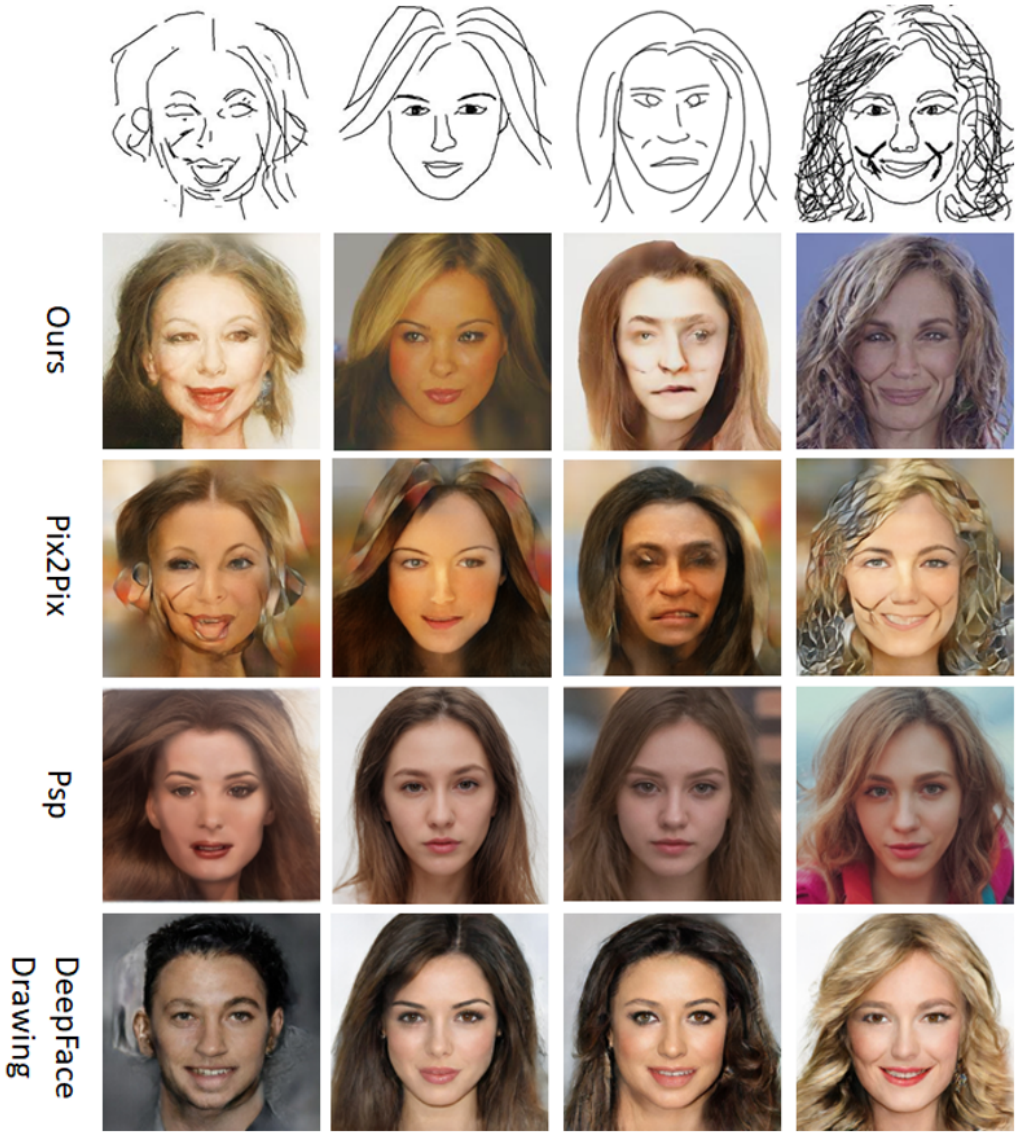

Figure 5.12: Less successful examples generated from the low-quality sketch inputs. Except for the second sketch from the right, the input sketches are from [7].

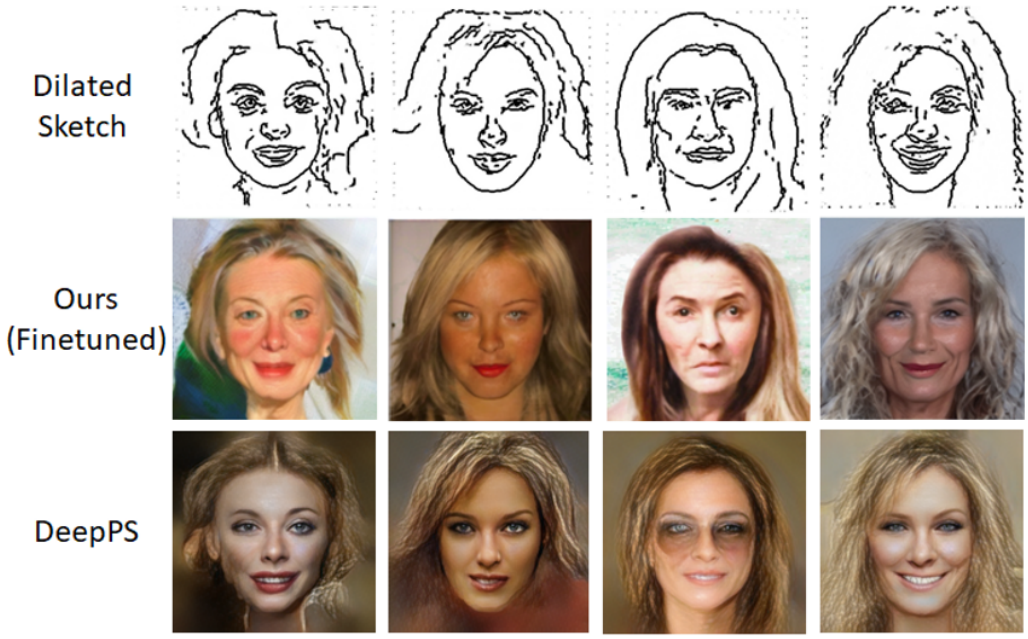

Figure 5.13: The results of finetuned model trained with dilated sketch data.

Chapter 6

Conclusion

6.1 Summary & Discussion

This thesis aims to incorporate the fundamental aspects of the traditional creative process, transitioning from *Ambiguous* to *Concrete* design intent, into the structure of modern generative models and design tools. In doing so, we analyze generative models through three performance metrics: Flexibility, Variability, and Precision. Using human image generation as an example, we divide the creative stages into three phases: the **Ideation Phase**, the **Processing Phase**, and **the Refinement Phase**. We assess the model performance requirements of creators at each stage and provide corresponding models and solutions based on these needs.

6.1.1 Evaluation of Models' Performance

To evaluate whether our models align with user needs in terms of performance across various creative stages, we conducted a questionnaire-based assessment. In this assessment, users were asked to rate the performance of the models as well as the clarity of design intentions at each stage. The rating was based on a five-point Likert scale. The experiment involved a total of six participants, including two Ph.D. students with a background in design, three graduate and Ph.D. students in computer science who lacked a design background and self-identified as having weak design skills (basic drawing abilities), and one Ph.D. student in computer science who, despite lacking formal training in design, had a passion for design and possessed some drawing skills.

As evident from Figure 6.1, the design intentions of all participants became progressively clearer through the three stages of the design process, evolving from initial ambiguity to eventual clarity. However, upon interviewing, the participant engaged in design-related work (ID: 2) and expressed that his work, even after being developed through our pipeline, still had room for further refinements, such as in the details of clothing, indicating that the design intention was not yet fully concrete.

Figure 6.2 demonstrates that, despite being a small-scale subjective scoring

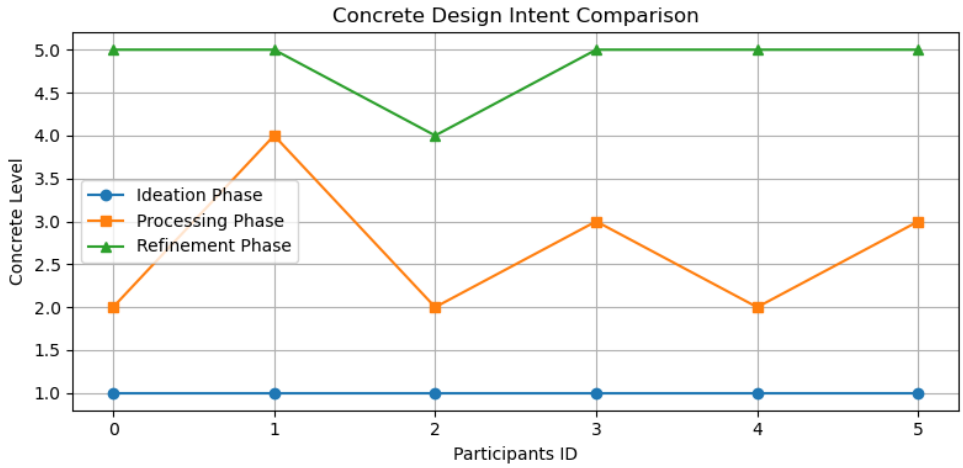

Figure 6.1: Participants' self-assessment scores for the clarity of their own design intentions at each stage.

experiment, the quantified results align with our initial hypotheses regarding the performance of models at different stages. In the presentation of results, the dashed line represents the median score, while the dotted lines indicate the first and third quartiles. In simple terms, the dotted lines show the scoring frequency that is second only to the most frequent score, excluding the median. Specifically, in the **Ideation Phase**, the model requires high flexibility; in the **Processing Phase**, high variability is needed; and in the **Refinement Phase**, high precision is crucial. One participant with a background in computer science suggested during an interview that, although the *HSI* model exhibits relatively lower flexibility, due to its specific input parameters like defined pose and clothing style, excessive flexibility and the resultant uncertainty might actually be counterproductive for design tasks in this phase. Indeed, given that the *HSI* model is designed to only allow inputs of pose and style (along with implicit text), it primarily caters to users whose design intentions are relatively clear (for instance, scoring 4 under the criteria of Figure 6.1). However, for users whose design intentions remain vague at this stage, a more flexible interaction approach is needed, such as allowing inputs from multiple modalities (e.g., explicit text prompts, clothing sketches, etc.).

Overall, our models are designed to reflect this dynamism and adaptability, enabling designers to interact with the medium conversationally and manage the creative process in line with their evolving ideas. Figure 6.3 illustrates the position of the proposed model within the performance space of the aforementioned three dimensions. Specifically, for creating human images, in the **Ideation Phase**, we offer *DualMotion*, a method that allows for the searching and editing of actions

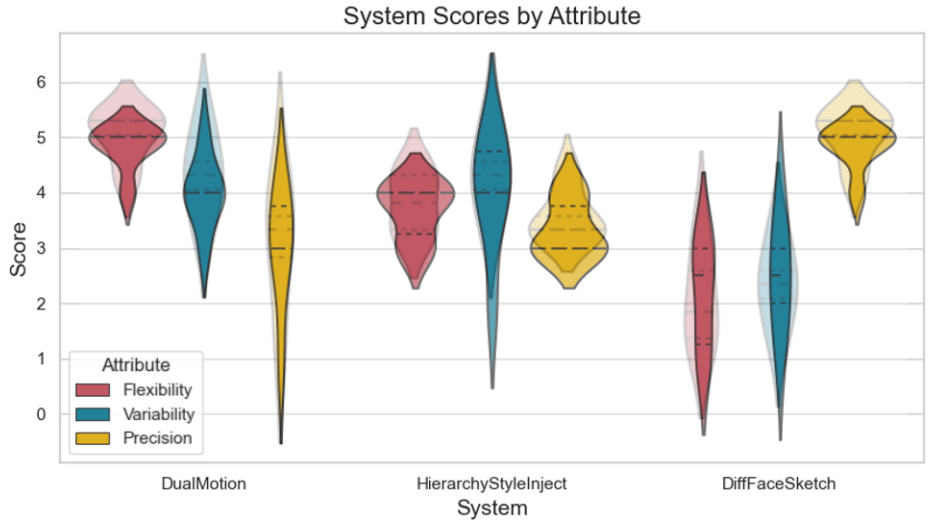

Figure 6.2: Participant Ratings for Each System's Model Performance in Flexibility, Precision, and Variability. (The lighter-colored legends represent scores from participants with a design background.).

by sketching motion trajectories. This gives users the ability to define poses from various angles using broad and abstract inputs, facilitating the exploration and capture of the desired pose. During the **Processing Phase**, we introduce a fine-tuning method for the large-scale pretrained text-to-image diffusion model *HSI* and train a pose and style-to-human generative model, which empowers users to create person images by freely altering poses and styles. This stage allows for the modification and trial of diverse combinations and outfits. Lastly, in the **Refinement Phase**, we present a High-Fidelity Face Image Synthesis diffusion model *DiffFaceSketch* that ensures a high level of consistency between the input and the output. This assures that users can fine-tune details with great precision while preserving the integrity of the areas that remain unedited.

Generative models facilitate human creative actions by understanding input conditions and generating valid samples. This process necessitates the provision of a shared knowledge space as a 'medium' between humans and models. As discussed in Section 2.2, in prior research, this 'medium' is often text, but textual semantics can vary in understanding even among humans. Particularly in image generation, while large-scale pretrained text-to-image models are powerful, they are not comprehensive and do not fully meet user needs at every stage of creation. Our proposed method leverages *HSI*, building upon a pretrained text-image model and fine-tuning it, allowing designers to guide image generation not just through text

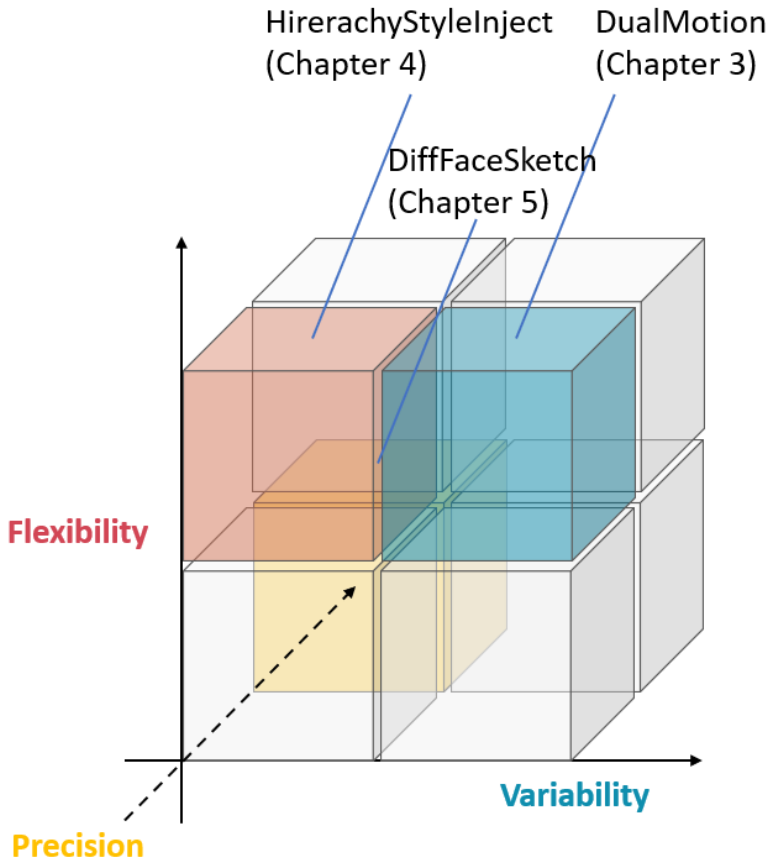

Figure 6.3: The summary of the proposed *DualMotion*, *HSI*, and *DiffFaceSketch* from the algorithms' performance within the three dimensions of Flexibility, Variability, and Precision.

prompts but also through image prompts. *DualMotion* and *DiffFaceSketch* move away from text-based priors, allowing users to input abstract motion trajectories (the former) or precise structural sketches (the latter) to produce results. We have demonstrated the effectiveness of these methods through empirical evidence and their suitability for different creative stages and model performance requirements.

6.1.2 Appropriate Target Users

To determine which user group is more suited for our proposed method, as mentioned in Section 6.1.1, we conducted surveys and interviews with users both with and without a design background. The lighter-shaded legend in Figure 6.2 represents the scoring results after excluding the three participants without a design

background. Interestingly, we found that these results more closely align with our initial hypotheses about the model's performance, particularly regarding the need for higher Flexibility in *DualMotion* and greater Precision in *DiffFaceSketch*. During the interviews, two of the participants appreciated the novelty and necessity of incorporating the clarity of design intentions into the interaction with the generative model. Another participant remarked that they had never considered the clarity of their design intentions during the creative process and that our definition of three design stages with assumed corresponding clarity levels served as a form of feedback for themselves (the user). This suggests that participants with a background in design are more suitable for using our designed pipeline.

To summarize, our research reimagines the function of generative models within the realm of creativity, moving from producing deterministic outputs to fostering a cooperative relationship between AI and artists. By implementing innovative methods along the spectrum from *Ambiguous* to *Concrete*, we create a reciprocal dynamic in which the model aligns with the artist's process instead of prescribing it. This transition from a deterministic to a heuristic approach represents a substantial advancement in the field of creative AI, nurturing a space where art and technology converge.

6.2 Remaining Challenges & Future Work

6.2.1 Continuous Solutions

In this thesis, we have explored the evolution of artistic creation in the era of data retrieval and large-scale generative models. We have argued that the real success of these technologies lies in their seamless integration into the artistic workflow, meeting the dynamic requirements of designers. Our methodology, highlighting diverse input modalities, interactivity, and adherence to the designer's workflow, suggests that the algorithms of generative models should be as flexible and responsive as the creative process they aim to emulate.

However, our proposed approach includes three separate models, and while we broadly categorize the creative process into three phases, this division can disrupt the continuity of creation. As illustrated in Figure 6.1, our three proposed methods are situated in three distinct locations within this three-dimensional performance space. Ideally, the model's performance should adapt and change in response to the user's design intent as it becomes clearer throughout the creative process. As we discussed in Section 1.4, user intent should progress from *Ambiguous* to *Concrete*, and in parallel, the performance of the ideal generative model should adapt smoothly and incrementally. We anticipate that future research will resolve

this challenge.

Figure 6.4: An example of HSI applied to the SD pretrained model Anything 4.0 [5] for a real-to-cartoon style transformation. The input is a real photo, and the output is an anime-stylized image.

Moreover, for certain specific tasks, such as performing an overall style transformation on generated images, the model's performance should ideally be situated at high variability and high precision. This means maintaining the image structure while transforming its style, as briefly mentioned in Section 4, where HSI is applied to the comic image fine-tuned *SD1.5* model *Anything4.0* [5], as shown in Figure 6.2. These scenarios could act as 'creative branches' outside the main creative process, and future research could provide a detailed analysis of these different task-specific model performance requirements.

6.2.2 Quantifying the Design Intent

Although this paper has hypothesized a quantitative relationship between the various performances of generative models and the clarity of design intent through discussions and analyses in Chapter 2, proposing different algorithms for different stages, and has empirically proven the effectiveness of these algorithms, it has not numerically quantified the clarity of design intent. To address the aforementioned

issue of model continuity, the best method would be to quantify the clarity of design intent and dynamically adjust the model's performance based on this value.

We believe that future research can gauge the completeness of work and quantify the clarity of design intent by having the model record and analyze user behavior during interaction. For instance, as discussed in Section 2.3, in the early stages of creation, the designer's actions are mostly 'creation', while in the later stages, the user's actions are primarily 'edition'. Therefore, we could determine the degree of completeness of the work and calculate the clarity of design intent by recording and analyzing whether the user's current behavior is more 'creation' or 'edition'.

6.2.3 Personalization & Customization

The complexity of human creativity cannot be fully replicated or anticipated by current generative models. The nuances of intuition, the serendipity of accidental strokes, and the depth of emotional intelligence inherent to human artists pose substantial challenges for computational models.

From my perspective, future work will involve:

- Developing more intuitive interfaces that require minimal technical expertise, thereby lowering barriers to entry for traditional designers and novices alike.
- Expanding the model's capabilities to comprehend and adapt to the nuances of personal style, thus offering a more bespoke creative experience.
- Extending our research to include collaborative AI, where the system can partake in the creative process as an active agent, capable of proposing ideas and alternatives that can inspire the human partner.

In summary, the dynamic interaction between generative models and human creativity presents a rich area for exploration, and our work has set the stage for a future where AI not only understands but also anticipates and elevates the human creative condition.

Acknowledgment

The journey to completing this thesis has been a transformative experience, marked by the unwavering guidance and encouragement of several remarkable mentors and the support of my peers and loved ones.

I am deeply indebted to my principal advisor, Professor Kazunori Miyata, for his invaluable mentorship, profound knowledge, and continuous support throughout my research endeavors. His keen insights and steadfast guidance have significantly shaped my academic journey and have been the compass by which I navigated this complex research landscape.

My academic voyage has been immensely enriched by the tutelage and expertise of Professor Haoran Xie, who has endured interest in the progression of my work and for his invaluable guidance and advice regarding the direction of my research. His attentiveness and astute counsel have been essential to my academic development.

The role of Professor Tsukasa Fukusato from Waseda University has been nothing short of foundational. Having guided me since the outset of my graduate studies, his mentorship has been a beacon of inspiration and academic excellence.

The professional stint at HUAWEI has been a significant milestone in my research career, and for that, I am eternally grateful to Bo Zheng, Erwin Wu, and my esteemed senior Naoya Iwamoto. They did not just mentor me in various projects but also provided strategic advice and support that were integral to the success of my research endeavors.

Also deserving of my gratitude is my co-advisor, Shogo Okada, whose expertise and thoughtful counsel have greatly contributed to my scholarly growth. His meticulous feedback and encouragement have been fundamental in refining my work to its current form.

I must also extend my gratitude to the JST SPRING, under Grant Number JPMJSP2102, for their financial support. This grant was pivotal in facilitating a focused and uninterrupted pursuit of my research objectives.

Lastly, this journey would have been a solitary one without the constant love and encouragement from my family, whose belief in my potential has never wavered. My friends and classmates have provided a network of camaraderie and support, for which I am ever so grateful. Their collective faith and encouragement have been a source of strength and motivation, driving me to strive for excellence.

References

[1] 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

[2] L. Weng, "What are diffusion models?" *lilianweng.github.io*, Jul 2021. [Online]. Available: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[3] Z. Lin, U. Ehsan, R. Agarwal, S. Dani, V. Vashishth, and M. O. Riedl, "Beyond prompts: Exploring the design space of mixed-initiative co-creativity systems," *International Conference on Computational Creativity*. [Online]. Available: https://par.nsf.gov/biblio/10434407

[4] "Sg161222 - realisticvision - v2.0 hugging face." [Online]. Available: https://huggingface.co/SG161222/Realistic_Vision_V2.0

[5] "Pastel-Mix [Stylized Anime Model] - Pastel-Mix [Pruned FP16] | Stable Diffusion Checkpoint | Civitai," 1 2023. [Online]. Available: https://civitai.com/models/5414/pastel-mix-stylized-anime-model

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2021, pp. 10 684–10 695.

[7] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deepfacedrawing: Deep generation of face images from sketches," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 72:1–72:16, 2020.

[8] 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

[9] "Adobe Illustrator - Industry-leading vector graphics software." [Online]. Available: https://www.adobe.com/products/illustrator.html

[10] “Official Adobe Photoshop - Leading AI photo design software.” [Online]. Available: https://www.adobe.com/products/photoshop.html

[11] “Creativity with AI — Deepart.io.” [Online]. Available: https://creativitywith.ai/deepartio/

[12] “OpenAI.” [Online]. Available: https://openai.com/

[13] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, “Improved variational inference with inverse autoregressive flow,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Paper.pdf

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, p. 139â□□144, oct 2020. [Online]. Available: https://doi.org/10.1145/3422622

[15] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[16] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:216078090

[17] J. Su and G. Wu, “f-vaes: Improve vaes with conditional flows,” 2018, cite arxiv:1809.05861. [Online]. Available: http://arxiv.org/abs/1809.05861

[18] L. Dinh, J. N. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *ArXiv*, vol. abs/1605.08803, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8768364

[19] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” *arXiv (Cornell University)*, 5 2021. [Online]. Available: http://arxiv.org/pdf/2105.05233.pdf

[20] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *arXiv (Cornell University)*, 12 2021. [Online]. Available: http://arxiv.org/abs/2112.10741

[21] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[22] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohenâ□□Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," *arXiv (Cornell University)*, 8 2022. [Online]. Available: http://arxiv.org/abs/2208.01618

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[24] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 1, pp. 1–13, 2022.

[25] 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," *arXiv preprint arXiv:2208.12242*, 2022.

[26] L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *IEEE International Conference on Computer Vision (ICCV)*, 2 2023. [Online]. Available: https://arxiv.org/abs/2302.05543

[27] 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," *arXiv preprint arXiv:2302.08453*, 2023.

[28] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022.

[29] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3588432.3591513

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

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[32] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[33] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[34] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 184–199.

[35] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2445–2454.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[37] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: conditioning method for denoising diffusion probabilistic models," *CoRR*, vol. abs/2108.02938, 2021. [Online]. Available: https://arxiv.org/abs/2108.02938

[38] C. Meng, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, "Sdedit: Image synthesis and editing with stochastic differential equations," *CoRR*, vol. abs/2108.01073, 2021. [Online]. Available: https://arxiv.org/abs/2108.01073

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

[40] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.

[41] D. Nettle, *Strong imagination: madness, creativity and human nature*, 1 2001. [Online]. Available: https://ci.nii.ac.jp/ncid/BA62470743

[42] M. Botella, F. Zenasni, and T. Lubart, "What are the stages of the creative process? what visual art students are saying." *Frontiers in Psychology*, vol. 9, 11 2018. [Online]. Available: https://doi.org/10.3389/fpsyg.2018.02266

[43] A. Vuichard, M. Botella, and I. C. Puozzo, "Creative Process and Multivariate Factors through a Creative Course "Keep Calm and Be Creative" ," *Journal of Intelligence*, vol. 11, no. 5, p. 83, 4 2023. [Online]. Available: https://doi.org/10.3390/jintelligence11050083

[44] H. International, "Creativity and human nature (What Wallace saw) - Hektoen International," 9 2022. [Online]. Available: https://hekint.org/2017/01/22/creativity-and-human-nature-what-wallace-saw/#:~:text=The%20ability%20to%20think%20creatively,plan%20and%20its%20effective

[45] P. Carruthers, S. Laurence, and S. Stich, *The Innate Mind, Volume 3*, 1 2008. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780195332834.001.0001

[46] A. Hertzmann, "Toward modeling creative processes for algorithmic painting," 2022.

[47] E. Protter, "Painters on painting (dover fine art, history of art)," http://dlib.net/, 2011.

[48] "Chuck Close | Pace Gallery," 1 2017. [Online]. Available: https://www.pacegallery.com/artists/chuck-close/

[49] D. Kahneman, *Thinking, fast and slow*, 1 2011. [Online]. Available: http://ci.nii.ac.jp/ncid/BB2184891X

[50] S. Yusuf, "From creativity to innovation," *Technology in Society*, vol. 31, no. 1, pp. 1–8, 2 2009. [Online]. Available: https://doi.org/10.1016/j.techsoc.2008.10.007

[51] S. Dadich, "Abstract：the art of design," 2017.

[52] I. Elrabbaaei, "Creativity spaces in graphic design: Pedagogical implications," *Academic Research International*, vol. 7, p. 4, 10 2018.

[53] A. Alvarez, J. Font, and J. Togelius, "Story designer: Towards a mixed-initiative tool to create narrative structures," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*, ser. FDG '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3555858.3555929

[54] A. J.A., "A pragmatic view of thematic analysis," *Qual Rep*, vol. 2, 11 1993.

[55] M. Agrawala, "Unpredictable Black Boxes are Terrible Interfaces," 3 2023. [Online]. Available: https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible

[56] R. K. Jones, T. Barton, X. Xu, K. Wang, E. Jiang, P. Guerrero, N. J. Mitra, and D. Ritchie, "Shapeassembly: Learning to generate programs for 3d shape structure synthesis," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, 2020.

[57] C.-H. Chiu, Y. Koyama, Y. Lai, T. Igarashi, and Y. Yue, "Human-in-the-loop differential subspace search in high-dimensional latent space," *ACM Transactions on Graphics*, vol. 39, no. 4, 8 2020. [Online]. Available: https://doi.org/10.1145/3386569.3392409

[58] Z. Huang, Y. Peng, T. Hibino, C. Z. Zhao, H. Xie, T. Fukusato, and K. Miyata, "dualface: Two-stage drawing guidance for freehand portrait sketching," *Computational Visual Media (CVMJ)*, vol. 8, pp. 64–77, 2022.

[59] Z. Huang, H. Xie, T. Fukusato, and K. Miyata, "Anifacedrawing: Anime portrait exploration during your sketching," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3588432.3591548

[60] "CMU graphics lab motion capture database," http://mocap.cs.cmu.edu./, 2022.

[61] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 7, pp. 1325–1339, 2014.

[62] M. G. Choi, K. Yang, T. Igarashi, J. Mitani, and J. Lee, "Retrieval and visualization of human motion data via stick figures," *Computer Graphics Forum*, vol. 31, no. 7, pp. 2057–2065, 2012.

[63] K. Brodt and M. Bessmeltsev, "Sketch2pose: Estimating a 3d character pose from a bitmap sketch," *ACM Transactions on Graphics*, vol. 41, no. 4, 7 2022.

[64] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[65] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[66] Y. Peng, Z. Huang, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "Sketch-based human motion retrieval via shadow guidance," in *Proceedings of Nicograph International (NicoInt'21)*. IEEE, 2021, pp. 42–45.

[67] Z. Hu, H. Xie, T. Fukusato, T. Sato, and T. Igarashi, "Sketch2vf: Sketch-based flow design with conditional generative adversarial network," *Computer Animation and Virtual Worlds (CAVW)*, vol. 30, no. 3–4, pp. e1889:1–e1889:11, 2019.

[68] R. H. Kazi, F. Chevalier, T. Grossman, S. Zhao, and G. Fitzmaurice, "Draco: Bringing life to illustrations with kinetic textures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. New York, NY, USA: ACM, 2014, pp. 351–360.

[69] J. Xing, R. H. Kazi, T. Grossman, L.-Y. Wei, J. Stam, and G. Fitzmaurice, "Energy-brushes: Interactive tools for illustrating stylized elemental dynamics," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST'16)*. New York, NY, USA: ACM, 2016, pp. 755–766.

[70] S. Zhang, Y. Guo, and Q. Gu, "Sketch2Model: View-aware 3D modeling from single free-hand sketches," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. Nashville, TN, USA: IEEE, 2021, pp. 6012–6021.

[71] M. Peng, J. Xing, and L.-Y. Wei, "Autocomplete 3D sculpting," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, pp. 132:1–132:15, 2018.

[72] R. H. Kazi, T. Grossman, N. Umetani, and G. Fitzmaurice, "Motion Amplifiers: sketching dynamic illustrations using the principles of 2D animation," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. New York, NY, USA: ACM, 2016, pp. 4599–4609.

[73] N. S. Willett, W. Li, J. Popovic, F. Berthouzoz, and A. Finkelstein, "Secondary motion for performed 2D animation," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST'17)*. New York, NY, USA: ACM, 2017, pp. 97–108.

[74] M. Dvorožňák, D. Sýkora, C. Curtis, B. Curless, O. Sorkine-Hornung, and D. Salesin, "Monster mash: A single-view approach to casual 3D modeling and animation," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 6, pp. 214:1–214:12, 2020.

[75] M. Guay, R. Ronfard, M. Gleicher, and M.-P. Cani, "Adding dynamics to sketch-based character animations," in *Proceedings of the Workshop on Sketch-Based Interfaces and Modeling (SBIM'15)*. Goslar, DEU: Eurographics Association, 2015, pp. 27–34.

[76] G. M. C. M. Guay Martin, Ronfard Remi, "Space-time sketching of character animation," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 118:1–118:10, 2015.

[77] M. Dontcheva, G. Yngve, and Z. Popović, "Layered acting for character animation," *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3, pp. 409–416, jul 2003. [Online]. Available: https://doi.org/10.1145/882262.882285

[78] E. S. L. Ho, H. P. H. Shum, Y.-m. Cheung, and P. C. Yuen, "Topology aware data-driven inverse kinematics," *Computer Graphics Forum*, vol. 32, no. 7, pp. 61–70, Oct 2013.

[79] N. Iwamoto, H. P. H. Shum, W. Asahina, and S. Morishima, "Automatic sign dance synthesis from gesture-based sign language," in *Proceedings of the 2019 International Conference on Motion, Interaction and Games (MIG'19)*. New York, NY, USA: ACM, 2019, pp. 18:1–18:9.

[80] B. Krüger, J. Tautges, A. Weber, and A. Zinke, "Fast local and global similarity searches in large motion capture databases," in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA'10)*. Goslar, DEU: Eurographics Association, 2010, pp. 1–10.

[81] M. Thorne, D. Burke, and M. van de Panne, "Motion doodles: An interface for sketching character motion," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 424–431, 2004.

[82] B. Choi, R. B. i Ribera, J. P. Lewis, Y. Seol, S. Hong, H. Eom, S. Jung, and J. Noh, "SketchiMo: Sketch-based motion editing for articulated characters," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 146:1–146:12, 2016.

[83] Q. L. Li, W. D. Geng, T. Yu, X. J. Shen, N. Lau, and G. Yu, "Motionmaster: Authoring and choreographing kung-fu motions by sketch drawings," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA'06)*. Goslar, DEU: Eurographics Association, 2006, pp. 233–241.

[84] Y. J. Lee, C. L. Zitnick, and M. F. Cohen, "Shadowdraw: Real-time user guidance for freehand drawing," *ACM Transactions on Graphics (ToG)*, vol. 30, no. 4, pp. 27:1–27:10, 2011.

[85] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*. CRC Press, 1996, pp. 207–212.

[86] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[87] D. of Health and H. Services, "System Usability Scale (SUS) | Usability.gov." [Online]. Available: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html#:~:text=Based%20on%20research%2C%20a%20SUS,in%20context%20about%20the%20process

[88] "SUS: The System Usability Scale - Trymata," 6 2023. [Online]. Available: https://trymata.com/learn/sus-system-usability-scale/#:~:text=Interpreting%20your%20System%20Usability%20Scale,website%E2%80%99s%20usability%20is%20about%20average

[89] J. Sauro, PhD, "5 ways to interpret a SUS Score - MeasuringU." [Online]. Available: https://measuringu.com/interpret-sus-score/#:~:text=1,to%20others%20in%20the%20database

[90] "Mixamo," https://www.mixamo.com/, 2022.

[91] Y. Peng, C. Zhao, Z. Huang, T. Fukusato, H. Xie, and K. Miyata, "Two-stage motion editing interface for character animation," in *The ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA'21)*. New York, NY, USA: ACM, 2021, pp. 3:1–3:2. [Online]. Available: https://doi.org/10.1145/3475946.3480960

[92] J. Hwang, I. H. Suh, G. Park, and T. Kwon, "Human character balancing motion generation based on a double inverted pendulum model," in *Proceedings of the Tenth International Conference on Motion in Games (MIG'17)*. New York, NY, USA: ACM, 2017, pp. 11:1–11:6.

[93] P. Esser and E. Sutter, "A variational u-net for conditional appearance and shape generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 8857–8866. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00923

[94] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, 2020, pp. 2287–2296.

[95] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[96] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[97] S. Y. Cheong, A. Mustafa, and A. Gilbert, "Upgpt: Universal diffusion model for person image generation, editing and pose transfer," October 2023, pp. 4173–4182.

[98] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[99] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.

[100] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[101] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 10 674–10 685. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042

[102] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete representation learning," *Neural Information Processing Systems*, vol. 30, pp. 6306–6315, 11 2017. [Online]. Available: http://papers.nips.cc/paper/7210-neural-discrete-representation-learning.pdf

[103] "stabilityai/sd-vae-ft-mse Â· Hugging Face." [Online]. Available: https://huggingface.co/stabilityai/sd-vae-ft-mse

[104] P. Esser, R. Rombach, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," *arXiv (Cornell University)*, 12 2020. [Online]. Available: http://arxiv.org/pdf/2012.09841.pdf

[105] C. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 5548–5557. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00559

[106] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 2337–2346.

[107] C. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, pp. 5548–5557.

[108] H. Caesar, J. R. R. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA: IEEE, 2018, pp. 1209–1218.

[109] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations (ICLR)*. Virtual Event, Austria: OpenReview.net, 2021, pp. 1–20. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[110] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8162–8171, 2021. [Online]. Available: http://proceedings.mlr.press/v139/nichol21a/nichol21a.pdf

[111] 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," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8748–8763, 2021. [Online]. Available: http://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[112] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: Fully convolutional networks for rough sketch cleanup," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 121:1–121:11, 2016.

[113] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Mastering sketching: Adversarial augmentation for structured prediction," *ACM Transactions on Graphics*, vol. 37, no. 1, pp. 11:1–11:13, 2018.

[114] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, pp. 2223–2232.

[115] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 8798–8807.

[116] C. Gao, Q. Liu, Q. Xu, J. Liu, L. Wang, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, pp. 5174–5183.

[117] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

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

[119] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, pp. 2849–2857.

[120] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Pastiche master: Exemplar-based high-resolution portrait style transfer," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022, pp. 7693–7702.

[121] M. PernuÅ¡, C. Fookes, V. Å truc, and S. DobriÅ¡ek, "Fice: Text-conditioned fashion image editing with guided gan inversion," 2023.

[122] U. Osahor and N. M. Nasrabadi, "Text-guided sketch-to-photo image synthesis," *IEEE Access*, vol. 10, pp. 98 278–98 289, 2022.

[123] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer, 2015, pp. 234–241.

[124] R. Yi, Y. Liu, Y. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 10 743–10 752.

[125] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "Mangagan: Unpaired photo-to-manga translation based on the methodology of manga drawing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2611–2619, 2021.

[126] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022, pp. 16 000–16 009.

[127] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM*

SIGGRAPH 2022 Conference Proceedings. New York, NY, USA: ACM, 2021, pp. 15:1–15:10.

[128] D. Horita, J. Yang, D. Chen, Y. Koyama, and K. Aizawa, "A structure-guided diffusion model for large-hole diverse image completion," *CoRR*, vol. arXiv:2211.10437, pp. 1–17, 2022.

[129] Adobe Systems Inc, "Photo to pencil sketch," https://www.adobe.com/creativecloud/photography/discover/photo-to-pencil-sketch.html, 2022.

[130] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 601–617.

[131] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *CoRR*, vol. abs/1506.03365, pp. 1–9, 2015. [Online]. Available: http://arxiv.org/abs/1506.03365

[132] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020, pp. 8188–8197.

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

Publications

[1] Yichen Peng, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, Kazunori Miyata, and Takeo Igarashi, *DualMotion: Global-to-Local Casual Motion Design for Character Animations*, IEICE TRANSACTIONS on Information and Systems, April 1st, 2023, Vol.E106-D, No.4, pp.459-468.

[2] Yichen Peng, Zhengyu Huang, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata, *Sketch-based human motion retrieval via shadow guidance*, IEEE Nicograph International (NicoInt), July 9th, 2021, pp.42-45. (Short Paper)

[3] Yichen Peng, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, Kazunori Miyata, and Takeo Igarashi, *Two-stage motion editing interface for character animation*, The ACM SIGGRAPH/Eurographics Symposium on Computer Animation, September 6th, 2021, pp.1-2. (Poster)

[4] Yichen Peng, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata, *Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis*, IEEE Access - The Multidisciplinary Open Access Journal, 2023 (Accepted)