

MotionCanvas: Generative Motion-Driven Interactive Art Experience

Qinxin Shu and Shuai Xu

DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2024

MASTER THESIS

SONY



MotionCanvas: Generative Motion-Driven Interactive Art Experience

Qinxin Shu and Shuai Xu



LUND
UNIVERSITY

MotionCanvas: Generative Motion-Driven Interactive Art Experience

Copyright © 2024 Qinxin Shu and Shuai Xu

Published by

Department of Design Sciences
Faculty of Engineering LTH, Lund University
Box 118
SE-221 00 LUND
Sweden

Supervisor(s): Günter Alce, gunter.alce@design.lth.se and

Sangxia Huang, sangxia.huang@sony.com

Examiner: Joakim Eriksson, joakim.eriksson@design.lth.se

Abstract

We introduce MotionCanvas, a spatial art creation tool that integrates Generative AI and motion capture technologies into artistic interactive experiences. Existing interactive platforms provide engaging environments and allow users to interact through touch or gestures. However, they are often limited by predefined interactive content and gestures, lacking the ability to respond to user movements comprehensively. We combine Generative AI models with motion capture technology to create novel forms of interactive art, with three primary objectives: 1) ensuring that the generated content aligns with the intended aesthetic, 2) advancing the possibilities for creative collaboration, and 3) minimizing latency within the interaction pipeline. The results of the user study confirm that our system can enhance user engagement and provides dynamic and immersive experiences.

Keywords: Interactive experience, Generative AI, Motion capture, Art creation

Sammanfattning

Vi presenterar MotionCanvas, en spatial konstskapande verktyg som integrerar generativ AI och rörelsekapningsteknologier i interaktiva konstupplevelser. Existerande interaktiva plattformar erbjuder engagerande miljöer och tillåter användare att interagera genom beröring eller gester. Men de är ofta begränsade av fördefinierad interaktiva innehåll och gester, och därmed saknar de möjligheten att reagera på användarens rörelser. Vi kombinerar generativa AI-modeller med spårningsteknologi för att skapa nya former av interaktiva konstverk, med tre primära mål: 1) säkerställa att genererat innehåll är i linje med avsett estetik, 2) främja möjligheterna för kreativt samarbete, och 3) minimera fördröjningen i interaktionen. Resultaten från användarstudien bekräftar att vårt system kan öka användarens engagemang och leverera dynamiska och immersiva upplevelser.

Nyckelord: Interaktiv Upplevelse, Generativ AI, Rörelsekapning, Konstskapande

Acknowledgements

We would like to express our sincere gratitude to our supervisor, Günter, for his invaluable guidance, support, and advice throughout this thesis. His insights and encouragement were instrumental in the successful completion of this thesis.

We also extend our heartfelt thanks to Sangxia for his help and constructive feedback in every stage of the thesis, from the idea exploration to implementation to user tests, which significantly contributed to the development of our work. Our appreciation also goes to the Sony Vision and AI Systems team for their comprehensive support. Special thanks go to Andrej, for his assistance and leadership, which were crucial to our progress. We are also grateful to our colleagues Sven, Leonardo, Nithin, Jonas and John-Henry, whose companionship during fika and enjoyment of semla provided much-needed moments of relaxation and happiness, helping us to release stress and maintain a positive outlook throughout this journey.

Lastly, we are immensely grateful to our friends and parents for their unwavering support, understanding, and encouragement throughout this endeavor. Their belief in us and their continuous encouragement have been invaluable.

This journey has truly been both pleasant and rewarding. Having the opportunity to learn, explore, and create within a field that deeply fascinates us has been a privilege. Being surrounded by talented individuals who were always ready to offer their assistance and insights, has been a true blessing. As this chapter comes to a close, the knowledge, experiences and memories we have gained will undoubtedly leave a lasting impact on our future endeavors, shaping our next steps in profound ways. Thank you all for being an invaluable part of our journey!

Lund, June 2024
Qinxin Shu and Shuai Xu

Table of contents

| | |
|--|----|
| List of acronyms and abbreviations | 9 |
| 1. Introduction | 10 |
| 1.1 Background | 10 |
| 1.2 Project Goals | 10 |
| 1.3 Research Questions | 11 |
| 1.4 Company | 11 |
| 1.5 Sustainable Development Goals | 11 |
| 1.6 Ethics | 12 |
| 1.7 Related Works | 13 |
| 2. Theory | 15 |
| 2.1 Motion Capture | 15 |
| 2.2 Generative Artificial Intelligence | 16 |
| 2.3 Interactive Experiences | 16 |
| 2.4 Image Processing | 17 |
| 2.5 Usability | 17 |
| 2.6 User Test | 18 |
| 3. Exploratory Study | 19 |
| 3.1 Scenario 1: Interactive Avatar Motion Generation | 19 |
| 3.1.1 Evaluation of Motion-to-Text and Text-to-Motion Models | 20 |
| 3.1.2 Compatibility Testing with Motion Capture Systems | 21 |
| 3.1.3 Evaluation of Large Language Models | 22 |
| 3.1.4 Summary | 23 |
| 3.2 Scenario 2: Interactive Art Content Generation with Motion | 23 |
| 3.2.1 Evaluation of Image Generation Models | 24 |
| 3.2.2 Image Post-Processing | 25 |
| 3.2.3 Summary | 26 |
| 4. MotionCanvas Prototype | 27 |
| 4.1 System Overview | 27 |
| 4.1.1 Hardware Settings | 27 |
| 4.1.2 Motion Capture System | 28 |
| 4.1.3 Software System | 29 |
| 4.2 Motion-to-Line Transformation Algorithm | 30 |
| 4.2.1 Data Collection and Processing | 30 |
| 4.2.2 Drawing Mechanism | 30 |
| 4.3 Flower Image Generation Pipeline | 32 |
| 4.3.1 Line Processing | 33 |
| 4.3.1.1 Line Smoothing Algorithm | 33 |
| 4.3.1.2 Geometric Flower Shape Addition Algorithm | 34 |
| 4.3.2 Stable Diffusion Model Configuration | 36 |

| | |
|--|----|
| 4.3.2.1 Model Construction Configuration | 36 |
| 4.3.2.2 Model Inference Configuration | 37 |
| 4.3.3 Task Queue Management | 38 |
| 4.4 User Interface Design | 38 |
| 5. User Evaluation | 39 |
| 5.1 Setup | 39 |
| 5.2 Participants | 40 |
| 5.3 Methods | 41 |
| 5.4 Procedure and Tasks | 41 |
| 5.4.1 Exploration Task | 42 |
| 5.4.2 Standardized Task | 43 |
| 5.4.3 Co-Creation Task | 44 |
| 5.5 Results | 45 |
| 5.5.1 Exploration Task Feedback | 45 |
| 5.5.2 Standardized Task Feedback | 46 |
| 5.5.3 Co-creation Task Feedback | 48 |
| 5.5.4 System Usability Scale | 49 |
| 6. Discussion | 51 |
| 6.1 Answer Research Question | 51 |
| 6.2 Limitation | 52 |
| 6.3 Future Work | 52 |
| 7. Conclusions | 54 |
| 7.1 Key Findings | 54 |
| 7.2 Contributions | 54 |
| 7.3 Recommendations | 55 |
| References | 56 |
| A.Appendix | 60 |
| A.1 Questionnaires for the user test | 60 |

List of acronyms and abbreviations

| | |
|----------|------------------------------------|
| AR | augmented reality |
| GANs | generative adversarial networks |
| GenAI | generative artificial intelligence |
| GPT | generative pre-trained transformer |
| LBE | location-based entertainment |
| LCM-LoRA | latent consistency model lora |
| LLMs | large language models |
| M2T | motion-to-text |
| SAM | segment anything model |
| SD v1.5 | stable diffusion v1.5 |
| SDGs | sustainable development goals |
| SDXL | stable diffusion xl |
| SMPL | skinned multi-Person linear |
| SUS | system usability scale |
| T2M | text-to-motion |
| VAEs | variational autoencoders |
| VR | virtual reality |
| XR | mixed reality |

1 Introduction

In this chapter, we introduce the background and put forth the motivation for this thesis project. As the foundation of the project, we also introduce the key technologies, including motion capture system and generative models.

1.1 Background

Interactive experience technologies have significantly evolved, providing immersive and engaging environments. Some platforms, such as immersive cubes and interactive screens, allow users to stay in a 2D/3D environment with visual content projected onto the surrounding walls or screens [15], [18]. Interactive screens enable users to interact with digital content through touch or gestures [58]. These technologies are widely used in museums, educational settings, and entertainment venues.

However, these systems often lack the ability to fully capture and respond to users' movements in real-time, limiting the level of interaction. Motion capture technology addresses this gap by precisely tracking users' bodies. Instead of relying on touch or predefined gestures, users can engage with the system through their motions, enhancing the immerse of the experience [43].

Another limitation of current interactive technologies is that the interactive content is often predefined, which limits its variety. Generative AI (GenAI) is an innovative tool that can generate diverse and unpredictable content, especially in artistic field, such as image, music and animation creations [12]. This capability enhances the diversity of interactive experiences and afford systems spontaneity and variability. The dynamic interactive content ensures that each interaction remains fresh and unpredictable, enhancing user engagement and fostering exploration within the interactive environment.

The combination of motion capture technology and GenAI enables the art to be interactive by inviting users to become co-creators rather than passive observers. The variability can attract a wider audience, as people return to see how the artwork changes with different interactions. Lastly, interactive art can serve educational purposes by engaging users in a more hands-on, immersive way, enhancing learning.

In our research, we aim to explore the potential of GenAI and motion capture in artistic interactive experiences. We seek to push the boundaries of artistic innovation and inspire new forms of interactive creation in the digital age.

1.2 Project Goals

The primary objective of this research is to investigate and develop a comprehensive pipeline for providing highly interactive experiences that bridge human motion with artistic creation, utilizing the synergy between GenAI models and motion capturing technology.

Specifically, this project focuses on transforming users' movements into dynamic inputs for the generative process in real-time. This capability sets the stage for the next phase, where GenAI models transform the inputs – ranging from simple sketches and complex gestures – into fine stylized paintings and artwork animations.

Moreover, a pivotal component of this research is to enable multi-person interactions, facilitating a co-creation experience. This aspect is designed to not only enhance the interactive experience but also to foster a sense of community and collaboration among participants.

1.3 Research Questions

Following the exploration of the evolving landscape of interactive experiences and the integration of GenAI with motion capture technology, several critical questions arise. These questions aim to address the technical gaps and enhance the practical applications of our research. Consequently, we have identified three primary research questions to guide our investigation:

- (1) **GenAI-Motion Integration:** How to combine GenAI models with the motion capture to create new form of interactive art experience?
- (2) **Interactive Content:** How to ensure that the generated content aligns with the overall style and visual aesthetic of an interactive experience?
- (3) **Low Latency:** How to minimize the potential latency during the interaction pipeline?

By addressing these research questions, our study intends to improve the technological foundation and user experience of interactive art platforms.

1.4 Company

This project benefits from collaboration with the Sony Nordic in Lund, Sweden, where we have access to state-of-the-art motion capture technology and expertise. Sony, a global leader in entertainment, technology, and innovation, provides invaluable support and resources that enhance the quality and capabilities of our interactive content project.

Sony's motion capture system offers precise tracking of human motion, optimized for real-time interactivity with minimal latency. It utilizes high-resolution cameras and marker-less tracking technology, supported by transformer-based neural network models, to accurately capture the subtle nuances of motion. By collaborating with Sony, we gain access to advanced facilities that enhance our ability to capture realistic human motion data, which is crucial for developing our interactive content experiences.

Sony's mission is to fill the world with emotion through the power of creativity and technology. This goal drives their commitment to supporting creators with innovative technologies that enhance storytelling and entertainment across various platforms.

Our collaboration with Sony, driven by their mission to fill the world with emotion through creativity and technology, is set to enhance Location-Based Entertainment (LBE) by making it more accessible and engaging for all skill levels. This project integrates GenAI to push the boundaries of LBE and expand the application of the motion capture system, increasing its capacity to create unique experiences. These efforts aim to redefine the intersection of art and technology, delivering compelling narratives that resonate across diverse audiences.

1.5 Sustainable Development Goals

The Sustainable Development Goals (SDGs), established by the United Nations in 2015, provide a framework for addressing global challenges and promoting sustainable development [21]. Our project aligns with Quality Education (Figure 1.1) and Industry, Innovation, and Infrastructure (Figure 1.2).

Goal 4: Quality Education



Figure 1.1 Quality Education.

By leveraging digital technologies, we empower individuals to explore new concepts, develop critical thinking skills, and foster creativity. Our interactive experiences are designed to be inclusive and accessible, ensuring that all people, including those with disabilities or other disadvantages, can benefit from the learning opportunities provided.

Goal 9: Industry, Innovation, and Infrastructure



Figure 1.2 Industry, Innovation, and Infrastructure.

Our project represents an innovative integration of motion capture systems, GenAI, and interactive content creation techniques. By leveraging these technologies, we drive advancements in digital content creation and interactive media. Through interdisciplinary collaboration and technological innovation, we contribute to the development of foster innovation and promote sustainable industrialization.

Our project intersects with SDGs. By harnessing the power of technology, creativity, and collaboration, we strive to contribute to a more inclusive, equitable, and sustainable future for all.

1.6 Ethics

The integration of GenAI and motion capture technology in interactive experiences raises important ethical considerations. Here are some potential ethics issues to consider:

Data Privacy and Consent

We adhere to data protection regulations and guidelines, ensuring that user data is collected, stored, and processed in a secure and responsible manner [53]. We are committed to ensuring that all participants

in the motion capture sessions provide informed consent regarding the collection and use of their data and clearly communicate how their motion data will be used and obtain explicit consent for its use in research and development.

Algorithmic Bias and Fairness

We recognize the potential for algorithmic bias to perpetuate or amplify existing inequalities and biases in society [52]. In the development of AI-driven components of our project, we assess and mitigate potential biases in the open-source diffusion AI model to ensure fairness and equity in its outputs.

Inclusivity and Accessibility

We strive to create interactive experiences that are inclusive and accessible to users of different backgrounds. This includes designing user interfaces and interactions that are intuitive and accessible to users of all abilities is essential for ensuring inclusivity and equitable access to interactive experiences [30]. We actively solicit feedback from diverse user groups to address barriers to participation.

Ethics considerations are central to ensure that our technologies and interactions are responsible, inclusive, and aligned with ethical principles. We are committed to upholding high standards of ethical conduct and promoting a user-friendly our systems.

1.7 Related Works

Interactive drawing projects that leverage motion capture technology and GenAI for dynamic content generation are at the forefront of creative exploration and innovation. Here, we examine relevant research and projects that inform and inspire the development of our own interactive drawing project.

Motion Capture

Motion capture technology has been widely used in interactive environments to track and interpret human movement for various applications. The study by Tsampounaris et al. [55] shows a whole-body interaction interface for exploring different visualizations of movement using real-time motion capture and 3D models. The primary objective is to apply these technologies in dance learning and improvisation within a creative, gamified context.

Generative Model

GenAI models have revolutionized content creation by enabling the generation of dynamic and realistic imagery. Li et al. [33] introduced GLIGEN, a novel approach to text-to-image generation that integrates captions and bounding boxes inputs. Zhou et al. [65] introduced InstructCTG, a novel framework for controlled text generation that addresses this challenge by incorporating constraints through natural language instructions.

Interaction Design

Real-time interaction and feedback are essential components of interactive projects, allowing users to engage with the experiences. Barmaki and Hughes' [5] work highlights the potential of real-time feedback in virtual rehearsal environments to enhance teacher preparation and nonverbal communication skills.

Multi-person interaction fosters collaboration, communication, and social presence among participants. Yi-Chun Du et al. [11] discovered that multi-person interaction led to higher stress levels compared to individual play. However, the collaborative aspect of multiplayer interaction resulted in more positive effects on learning outcomes.

Collaborative Drawing

Collaborative drawing platforms have emerged as popular tools for artistic collaboration and expression among multiple users. In a study conducted by Lyon et al. [37], the authors investigated the impact of a collaborative drawing module themed on the human body. One significant finding was the development of "critical looking" skills through the drawing exercises, which were likened to processes involved in

clinical examination and diagnosis.

By drawing upon the insights and advancements of these related works, the integration of motion capture technology with GenAI holds immense potential for creating interactive drawing experiences. Our project aims to push the boundaries of interactive experiences by seamlessly integrating motion capture technology, GenAI, and real-time interaction to create dynamic and engaging content that inspires creativity and collaboration among multi users.

2 Theory

In this chapter, we provide an overview of the theoretical foundations and technologies for our project. We explain concepts related to motion capture systems, GenAI, interactive experiences, image processing, usability, and user testing methodologies.

2.1 Motion Capture

Motion capture technology are used within diverse industries and applications. In the film, television, and video game industries, motion capture is used to animate digital characters and create realistic computer-generated imagery effects [29]. In sports science, rehabilitation, and ergonomics research, motion capture enables precise analysis of human movement patterns, biomechanical efficiency, and injury prevention [43].

These systems typically consist of multiple cameras or sensors which record the positions of markers from different directions. By triangulating the positions of markers in 3D space, the subject's movement will be reconstructed in real-time or offline, allowing for accurate playback and analysis [2].

Motion capture systems come in various configurations and implementations, as presented in Figure 2.1, each suited to different applications and environments.

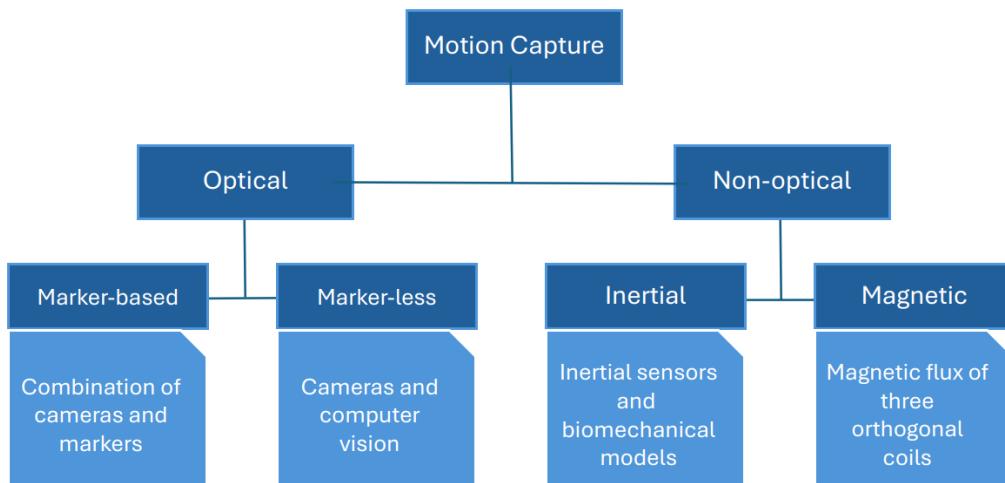


Figure 2.1 Motion Capture Methods.

Optical motion capture systems are camera-based techniques, comprising both marker-based and marker-less methodologies [60]. Marker-based systems use cameras equipped with infrared or visible light sensors to track the movement of reflective markers placed on a subject's body. Marker-less systems rely on computer vision to track and analyze the movement of objects. Machine learning also be employed to improve the accuracy and robustness of marker-less motion capture systems [16].

Non-optical motion capture systems utilize various technologies such as inertial sensors, magnetic fields, mechanical models and strain gauges or flex sensors et al. Inertial motion capture systems [57] use inertial measurement units to measure the acceleration, angular velocity, and orientation of body segments.

Magnetic motion capture systems rely on magnetic fields generated by sensors and transmitters to track the positions of markers [42].

For our interactive experience project, motion capture serves as a critical component for reconstructing users' movements. This integration enable users to use their own body movements as input, which enhances users' sense of participation.

2.2 Generative Artificial Intelligence

GenAI is a class of machine learning models that can generate new content based on existing data it was trained on. These models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), transformer architectures such as Generative Pre-trained Transformer (GPT) models and diffusion models. They have been used in various areas, including image generation, text generation, music composition and animations [7].

GANs are a class of generative models which train two neural networks simultaneously: a generator and a discriminator [17]. The generator generates synthetic data samples from random noise input. The discriminator evaluates the generated samples and tries to distinguish them from real data. GANs excel at generating high-quality, realistic samples across various domains, including images, text, and sound.

VAEs are probabilistic generative models which consist of an encoder and a decoder, trained to jointly learn a compressed representation of input data [26]. The encoder maps input data to a lower-dimensional latent space, capturing its underlying distribution. The decoder reconstructs the input data from samples drawn from the latent space, generating synthetic samples. VAEs provide a probabilistic framework for learning interpretable latent representations of data.

Transformers have revolutionized natural language processing and text generation tasks [25]. By leveraging self-attention mechanisms and large-scale pre-training, transformers achieve remarkable performance in capturing complex patterns and generating contextually coherent text.

Diffusion models are a class of generative models which aim at capturing high-dimensional data distributions. With their ability to accurately estimate likelihoods and generate high-quality samples, diffusion models have emerged as an approach for image synthesis and denoising [10].

One of the key strengths of GenAI is producing content that goes beyond what is explicitly programmed or defined. For our project, these novel content enables more dynamic and diverse responses for users.

2.3 Interactive Experiences

Interactive experience refers to the dynamic interaction between users and the digital environment. It involves the engagement between the user's actions or inputs and the system's responses. Immersive interactive experiences usually resonate with multiple senses of the users, such as visual, auditory, tactile and even olfactory stimulation, which helps to enhance immersion [14]. In immersive interactive experiences, users are active participants who interact with the digital environment.

Interactive experiences evoke emotional responses in users, such as excitement, curiosity and satisfaction. They engage users through engaging storytelling, visuals, and other sensory feedback to generate emotional responses [39], driving deeper engagement and promoting further interaction.

Virtual Reality(VR), Augmented Reality (AR) and Mixed Reality (XR) encompass a range of experiences that blend elements of the physical and digital worlds. VR allows users to interact with virtual objects [3]. AR overlays digital information onto a user's real-world environment [8]. XR maintain awareness of the physical environment [48].

In settings like museums or art galleries, large-scale interactions combine physical sensors, projection mapping and audio-visual effects to create an immersive and participatory experience for viewers. These technologies provide users with active interactions and focus on users' emotions, experiences and cognition [38].

For our project, we focus on delivering immersive and socially interactive experiences by timely and meaningful feedback and co-creation between users. Through these efforts, we aim to create dynamic and inclusive environments and a user-centered design.

2.4 Image Processing

Image processing is a technology that transform images into digital forms and perform certain operations to get some useful information [51]. It plays a crucial role in our project, enabling various functions such as background removal, handling depth information, and color manipulation. These techniques help transform the images generated by GenAI into visually appealing and interactive content.

Background removal involves isolating the foreground objects from the background in an image or video. This technique is essential for creating appealing compositions and removing distractions. Techniques such as chroma keying [4] and foreground segmentation [20] are employed to achieve this.

Image depth estimation is an image processing technique that analyze the spatial information associated with each pixel in an image, indicating the distance of objects from the camera or sensor. It provides a three-dimensional representation of the scene, allowing for accurate depth perception and spatial understanding. Image depth is commonly used in various applications, including AR, VR, robotics, and medical imaging [22].

Color manipulation is the process of modifying or transforming the color and tone of images or graphical elements. It allows adjusting the color palette, contrast and brightness, improving the appearance of content for enhanced visual appeal [40]. Whether used for enhancing realism, creating mood, or adding stylistic flair, color manipulation is an essential tool in the creative work.

Image processing techniques are integral to the realization of our project's objectives, enabling background removal, depth information extraction, and color manipulation to enhance the visual quality of the interactive content. By leveraging these techniques with motion capture data and GenAI models, we can create immersive and engaging experiences.

2.5 Usability

Usability refers to the extent to which a system enables users to achieve their goals with effectiveness, efficiency, and satisfaction. It includes learnability, ease of use, memorability, error prevention and recovery, and user satisfaction [50]. In our project, usability extends traditional user interfaces to encompass the entire user experience within the interactive motion capture environment.

Learnability, defined as the ease with which users can understand and operate a system, is crucial for enabling users to quickly grasp the system's functionalities and controls. Ease of use ensures that users can efficiently utilize the system to its fullest potential. Memorability, referring to the ease with which users can recall how to use the system over time, is essential for maintaining sustained usability and user satisfaction.

Usability-focused design reduces errors by providing clear feedback, intuitive controls, and error-prevention mechanisms, which allow error recovery, minimizes user frustration, and maintains workflow continuity [41]. These principles contribute to overall user satisfaction by promoting positive feedback and encouraging continued engagement with the system.

2.6 User Test

User testing is a key aspect of the development process for interactive systems, ensuring usability, functionality, and overall user satisfaction. In multi-user interaction systems, user testing becomes more crucial due to the complexities involved in coordinating interactions among multiple users.

Gathering user feedback is crucial for evaluating the system's performance. Clear objectives must be established, defining specific goals and hypotheses for the testing process. Tasks and scenarios presented to users during testing should mirror real-world usage scenarios, enabling a realistic assessment of system usability [44]. Maintaining unbiased observation throughout the testing process is also critical principles to adhere to.

Interviews and questionnaires are methods for getting qualitative and quantitative insights from users about their experiences, preferences, and suggestions for improvement [56]. By asking open-ended questions, we can get user preferences, challenges, and areas of interest that may not be apparent through other methods.

The System Usability Scale (SUS) is a widely used standardized questionnaire for the assessment of perceived usability [31]. SUS consists of ten statements and each of them is rated on a 5-point scale, ranging from strongly disagree to strongly agree. Odd-numbered questions are positive statements and even-numbered questions are negative statements. Participants should evaluate their level of agreement with each statement based on their experience with the system. We will use SUS to assess the usability of our system through a series of standardized questions. After analysing feedback from users, we can iterate system, leading to a more user-centered and effective solution.

3 Exploratory Study

In this chapter, we detail the exploratory study that assesses the feasibility of our interactive content creation pipeline. We focus on identifying the most suitable models for effective implementation.

In order to determine the feasibility of our interactive content creation pipeline and identify the necessary phases for its development, we conducted an exploratory study. This phase involved initial testing, evaluation, and iteration of various models and techniques across different scenarios. The aim was to develop a clear and feasible plan for structuring the system, while focusing on finding the most effective ways to integrate GenAI into our pipeline. All model evaluations and tests were conducted on a PC connected to a server equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of GDDR6X RAM. This hardware setup was chosen not only for its high computational power and graphics processing capability but also to ensure a consistent running base across this and subsequent prototype phases of the project.

Through these efforts, we sought to identify viable solutions and refine our approach, ultimately leading to a concrete idea of empowering the creation of interactive content with GenAI. This process ensured that our design would be practical and effective, setting a solid foundation for the next phases of development.

3.1 Scenario 1: Interactive Avatar Motion Generation

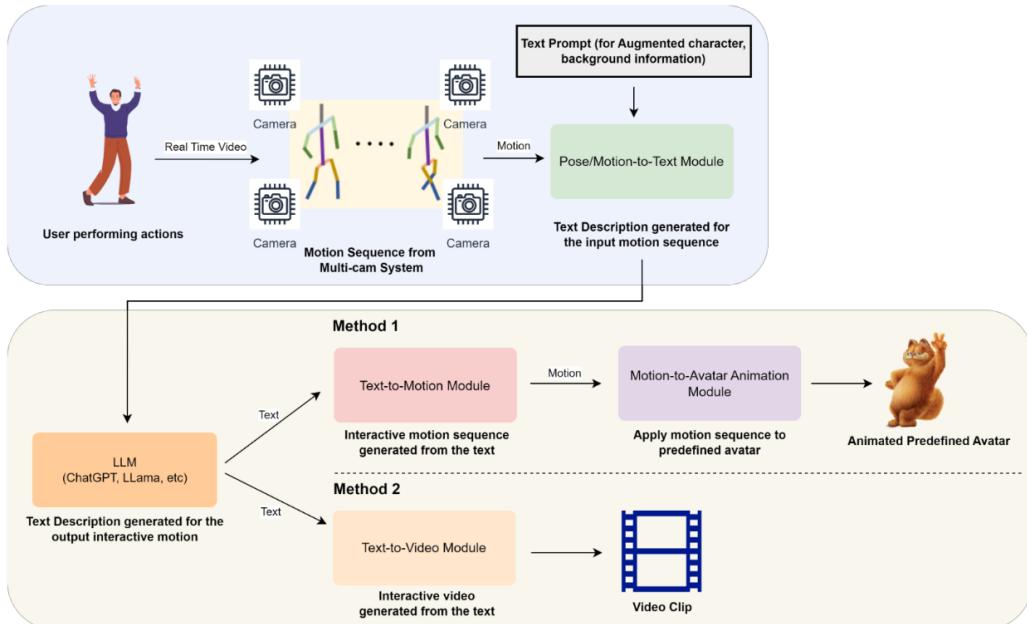


Figure 3.1 Pipeline Example of Scenario 1.

We started from investigating the relevant models and the feasibility of our first idea: Generating interactive avatar motion based on the user's input motions. Figure 3.1 shows an example pipeline of this process. This process involves understanding user's input motion and generating corresponding

interactive motions in return. The key models in this process would be Motion-Language Models for the conversion between corresponding motions and text descriptions, and the Large Language models (LLMs) for giving the motion text descriptions based on the received input text motion description.

3.1.1 Evaluation of Motion-to-Text and Text-to-Motion Models

The capability to perform both Motion-to-Text (M2T) and Text-to-Motion (T2M) tasks is a critical requirement for our pipeline. The T2M task involves generating human motion sequences from textual input, while the M2T task entails creating text descriptions based on human motion sequences. There are more models dedicated to T2M tasks than M2T tasks. Examples of such models include MotionCLIP [54], T2m-gpt [61], TM2T [19], which have been used for tasks like motion generation and motion captioning. Among these models, MotionGPT [23] provides a unified motion-language solution capable of performing both M2T and T2M tasks, with competitive results across a range of motion tasks. Therefore, MotionGPT is our primary choice for these tasks.

To assess the suitability of the model, we focused on two key aspects: the quality of generated outputs and the time required for generation.

For the T2M task, we prepared two text prompts, one with a simple description containing one action, and one with a more complex description, as presented in Figure 3.2. We observed that MotionGPT produced satisfactory results when provided with simple prompts, whereas its performance declined with complex prompts. Additionally, the generation of a motion typically required between 30 to 70 seconds.

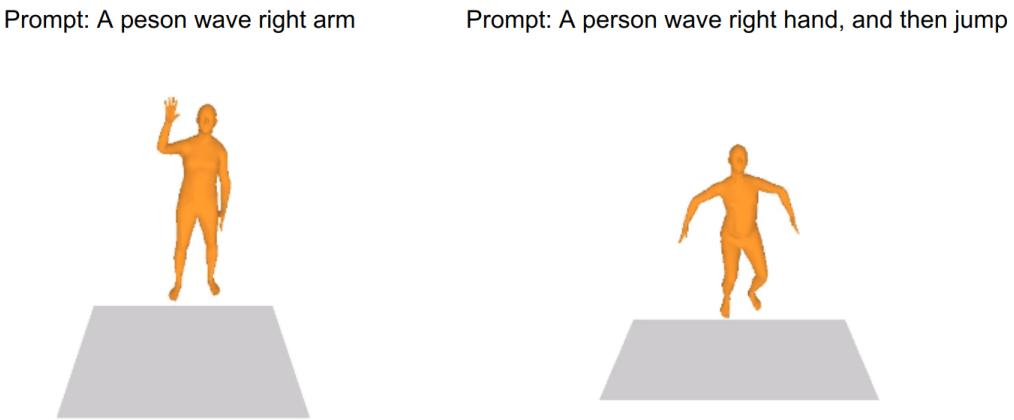


Figure 3.2 Results of Text-to-Motion task with MotionGPT.

Similarly, for the M2T task, we prepared two motion sequences, as presented in Figure 3.3. The MotionGPT showed good performance in explaining motion; however, it often required more than one minute to generate the explanations.

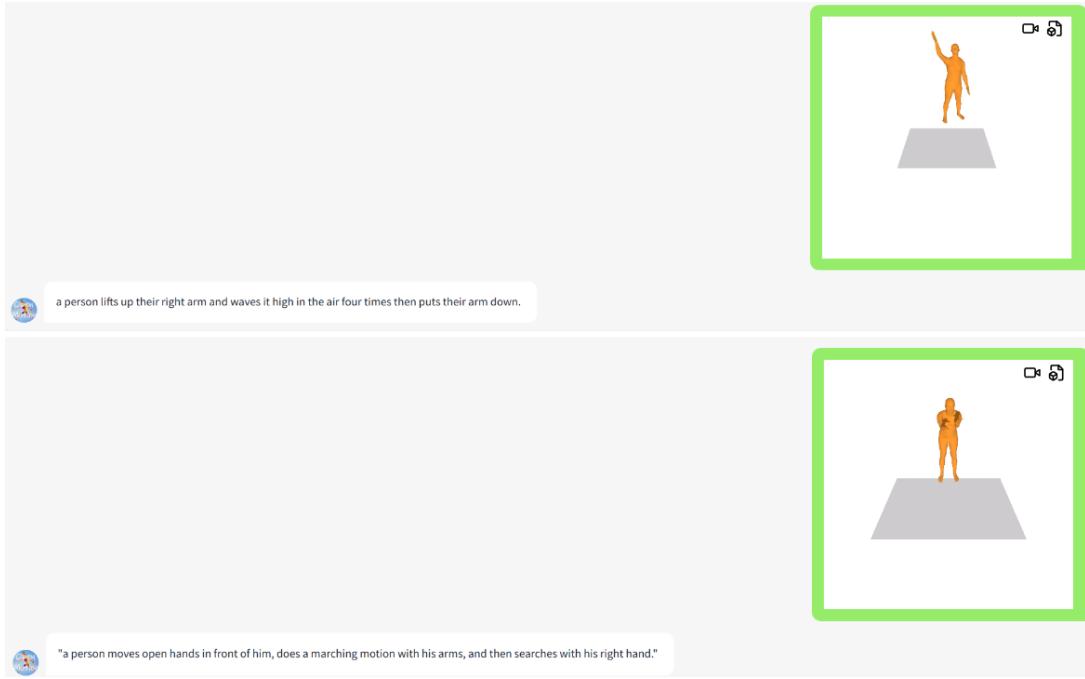


Figure 3.3 Results of Motion-to-Text task with MotionGPT.

3.1.2 Compatibility Testing with Motion Capture Systems

Human motion data can be represented in a variety of formats, depending on the specific requirements of the motion capture system. For our system, the motion capture data must be compatible with a human skeleton consisting of 14 body joints within a 3D coordinate system, forming a pose vector of 14×3 scalar values. A visualization of the skeleton model of our motion capture system is shown in Figure 3.4.

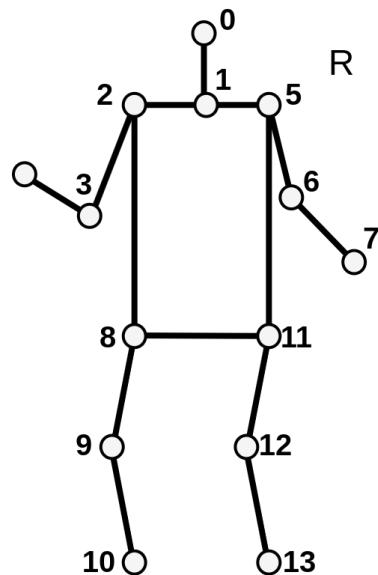


Figure 3.4 Skeleton Model of Our Motion Capture System.

Throughout our research, we discovered that leading Motion-Language Models, including MotionGPT, use the Skinned Multi-Person Linear (SMPL) model [35] as the foundation for encoding human subjects.

The SMPL model employs two types of parameters: shape and pose. The shape parameter is a vector consisting of 10 scalar values, representing the degree of expansion or contraction of a human subject. The pose parameter comprises a 24-joint hierarchy based on a kinematic tree structure that keeps the parent relation for each joint. The pose vector has 24×3 scalar values, where each joint's rotation is encoded as an arbitrary 3D vector in the axis-angle rotation representation. Figure 3.5 shows the visualization of the SMPL human model.

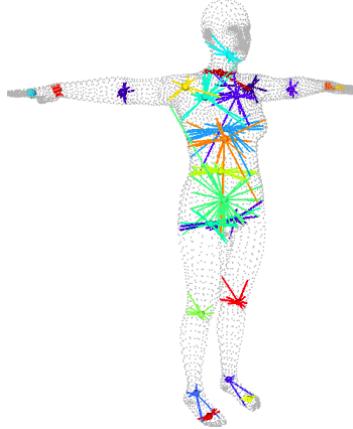


Figure 3.5 SMPL Human Model [35].

To convert our motion data into SMPL format, we explored several open-source solutions, such as Pose-to-SMPL [32], which fits SMPL parameters from 3D-pose datasets containing key points of the human body, and aitviewer [24], which provides tools for visualizing SMPL human data and its underlying skeletal structure. However, it proved challenging to convert between these two formats due to fundamental differences in joint definitions, parameterization, and the skinning methods. SMPL uses a 24-joint hierarchy based on a kinematic tree, while our system employs a simpler 14-joint structure. Additionally, SMPL's Linear Blend Skinning requires specific weights for each vertex, further complicating the conversion process.

3.1.3 Evaluation of Large Language Models

The generation of interactive motion sequences relies heavily on the creation of reliable response motion textual descriptions. This step is crucial in linking M2T and T2M modules within the process. LLMs have proven their capability in comprehending and generating contextually accurate responses, making them a suitable choice for this task. In particular, we examined GPT-4 [1] and LLaVA-1.5-7B [34] to evaluate their performance in generating textual responses for interactive motions.

To assess the models, we prepared ten sets of motion textual descriptions and asked them to generate a corresponding response. To improve consistency, we provided a guiding context prior to posing the questions, as follows:

"To build an interactive motion generation model, you need to create corresponding motion descriptions in response to given inputs. In this setup, there are two entities, A and B. When given a motion description for entity A, generate an appropriate response for entity B. Ensure the response contains at most two actions and is simple in nature. The actions should only contain the ones using arms and legs. Here's an example:

Input: A waves his right hand to B.

Output: B waves his right hand back to A. "

The responses from each model for the given input are summarized in Table 3.1. From the analysis, we can infer that both models produce very similar responses to given inputs. They mainly replicate the

input actions with little variation. For example, both models responded to "A waves his right hand to B" with "B waves his right hand back to A." This pattern was consistent across other inputs, indicating a lack of diversity and creativity in their outputs.

Table 3.1 Comparison of outputs from GPT-4 and LLaVA-1.6-34B

| Input | Output from GPT-4 | Output from LLaVA-1.6-34B |
|------------------------------|-----------------------------------|-----------------------------------|
| A waves his right hand to B. | B waves his right hand back to A. | B waves his right hand back to A. |
| A raises his left hand. | B raises his left hand to A. | B raises his left hand. |
| A taps his left foot. | B taps his right foot. | B taps his left foot. |
| A claps his hands. | B claps his hands. | B claps his hand. |
| A jumps up and down. | B jumps up and down. | B jumps up and down. |

3.1.4 Summary

Based on the evaluation of Motion-Language models, including data compatibility and the LLMs' ability to generate interactive motion textual descriptions, achieving real-time interactive avatar motion generation poses significant challenges. Although these models can generate reasonable motion sequences from textual input, the generation process takes more than 30 seconds, which is longer than expected for real-time applications. Additionally, the content and quality of the generated sequences lack the desired level of controllability. Compatibility issues between various data formats add to these challenges.

Given these limitations, we decided not to proceed with developing a real-time interactive motion generation pipeline using these models. The findings highlight the need for low-latency solutions that offer a balanced approach to content controllability and diversity. These two criteria are crucial for creating a reliable pipeline capable of supporting real-time interactive content generation through GenAI.

3.2 Scenario 2: Interactive Art Content Generation with Motion

Based on our earlier exploration, we began examining models capable of generating high-quality content with low latency, while allowing for controllability to adapt to various use cases. Among all types of generative models, image generation models lead the way, renowned for their ability to produce high-quality images with remarkable stability and diversity. Research indicates that text-to-image GenAI significantly enhances creative production [64] and can broaden the overall diversity of artistic outputs with a low barrier to entry [12]. Meanwhile, human motion, a natural form of interaction with the physical world [13], has been increasingly integrated into various interactive experiences. Given these insights, we developed the concept of generating interactive art content using motion as the input, with an image generation model allowing for the creation of fine-styled art.

To effectively demonstrate this concept, we decided to limit the scope of our project to a single theme: flower images. Focusing on a specific theme helps prove the model's effectiveness and is representative of real-world scenarios, which often center around specific themes such as promoting a character or hosting an art exhibition. We chose flowers as our theme because they are a common subject in art and are familiar to many people. This choice allows us to demonstrate the capabilities of our model in a familiar context, making the technology's benefits more relatable and understandable.

To evaluate the feasibility of this concept, we examined relevant models, focusing on factors including generation quality, generation time, model controllability, and required image processing methods.

3.2.1 Evaluation of Image Generation Models

For image generation models, we mainly looked into diffusion models, which are generative models that have been gaining state-of-the-art performance regarding the image generation task [10]. Diffusion models operate by initially introducing noise to data, subsequently reversing this process to reconstruct the data from its noisy state. The training process involves iteratively estimating the score function during each denoising step. This score function acts as a gradient that guides the model towards higher probability data points with reduced noise, as depicted in Figure 3.6 [59].

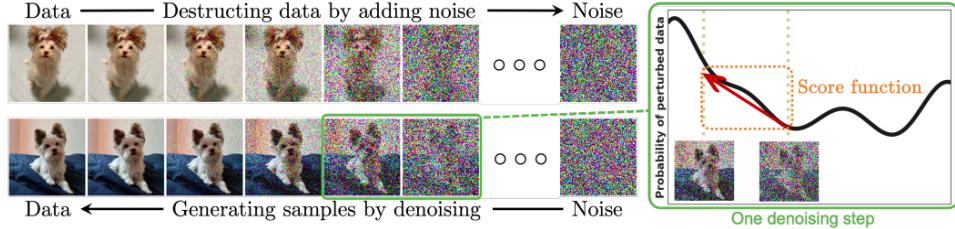


Figure 3.6 Overview of Diffusion Models' Training and Generation Process [59].

We chose to use the Diffusers library [45] on Hugging Face Hub¹ as our model research foundation due to its popularity, open-source accessibility, and robust support from a growing community of developers and researchers.

StreamDiffusion

Since the interactive experience requires real-time feedback from user input, we first looked into the StreamDiffusion model [28], which is a real-time diffusion pipeline developed to generate interactive and realistic images.

We tested the model on Image-to-Image tasks with additional text prompt to generate images. Figure 3.7 shows one example of the generated results. Regarding the image quality, however, we encountered challenges in maintaining content coherence with user inputs. The StreamDiffusion model attempts to represent all elements of the prompt in a single rendering, despite the user only drawing a partial picture. This disrupted the visual and conceptual continuity between the user's input and the generated content, detracting from the overall effectiveness of the experience.



Figure 3.7 Result of Image-to-Image Task with StreamDiffusion: Left - Input image, Right - Output image

Stable Diffusion and ControlNet

Building on the aforementioned points, we decided to seek models that offer better controllability while

¹ <https://huggingface.co/>

maintaining high generation quality. For diffusion models, one major method to enable additional input conditions is by incorporating the ControlNet model [62]. ControlNet is neural network model for controlling image diffusion models by conditioning the model with an additional input image, allowing for conditioning inputs such as line scribbles, depth information, and human pose inputs.

To integrate ControlNet, we investigated the two most-downloaded diffusion-based models on the Hugging Face platform, which are Stable Diffusion v1.5 (SD v1.5) [49] and Stable Diffusion XL (SDXL) [46]. Both models are known for generating high-quality images at resolutions of 512×512 and 1024×1024 , respectively, and are compatible with various model extensions, including ControlNet.

Considering the use of user motions as inputs to generate flower images, and the necessity for the generation results to align with these inputs, we investigated three types of conditioning inputs deemed suitable for our project. These are 1) Inpainting, which employs an additional mask image to indicate the desired area for image generation, ensuring that the generated image blends seamlessly with the surrounding areas; 2) Canny edge, which uses a monochrome image with white edges on a black background to guide the shape of the generated images; and 3) Depth conditioning, which utilizes a grayscale image carrying depth information to influence the generation results.

We tested these different conditioning inputs using both the SD v1.5 model and the SDXL model. From the test results, we found that, in terms of controllability, the inpainting condition occasionally fails to capture finer details, such as thinner strokes. On the other hand, both the canny edge and depth conditions effectively guide the generation results. The SDXL model provides better support for higher resolutions, notably at 1024×1024 pixels. For generation time, the average processing time is 1.5-2.5 seconds for the SD v1.5 model and 6-7 seconds for the SDXL model with standard 50 denoising steps. While the SDXL model excels in producing high-quality images and aligning effectively with specified prompt conditions through the ControlNet model, reducing the generation time remains a significant challenge to achieve more efficient interactive experiences. Despite the challenge, the choice of SDXL is favored particularly for its scalability and its ability to maintain high resolution, which are crucial for large interactive screens in LBE settings.

LCM-LoRA

To address this challenge of lengthy generation times, we explore the potential of Latent Consistency Model LoRA (LCM-LoRA) [36]. LCM-LoRA is an acceleration module that can be plugged into various Stable Diffusion models, and is known for its efficiency improvements in reducing the number of inference steps to only between 2-8 steps, which could lead to a shorter generation time and significantly enhance the responsiveness of our interactive systems. In comparison, the standard number of inference steps of SDXL is 50. To evaluate the extent to which LCM-LoRA can reduce generation time while maintaining image quality, we integrated it into the existing SDXL model and conducted tests at 4 and 8 inference steps. For comparison, we also ran the SDXL model at the same inference steps without the LCM-LoRA module, as well as at the default setting of 50 steps.

From the results, we observed that the LCM-LoRA module enables the production of high-quality images with shorter generation time. Although the images generated with 50 steps exhibit the finest details, the generation time is 6-7 seconds which is longer than then 0.8-second average when using 4 inference steps with LCM-LoRA. Considering the critical need for real-time interaction in our system, opting for a reasonably high image quality with significantly reduced generation time presents a more viable solution.

3.2.2 Image Post-Processing

From the previous experiments, we observed that the generated images included background colors, which could detract from the focus on the main subjects in our interactive art creation experience. Removing these backgrounds is essential to enhance visual clarity and viewer engagement with the art pieces. Therefore, it is crucial to identify an effective method for background removal that preserves the

main subject and its intricate details, such as branches and the varied shapes of flowers.

Based on public reviews and benchmark rankings of model performance in image segmentation tasks, we selected three models believed to be well-suited for our needs: 1) Segment Anything model (SAM) [27], 2) IS-Net [6], and 3) BiRefNet [63].

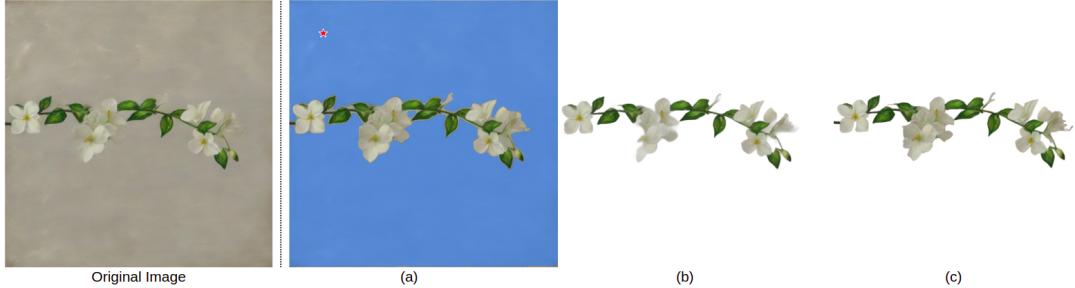


Figure 3.8 Image Segmentation Results: (a) SAM, (b) IS-Net, (c) BiRefNet.

Evaluating the results, we noted differences in removal quality, generation time, and ease of integration among the models. Figure 3.8 shows the results of 3 different image segmentation models. The SAM showed promising results in terms of removal quality when using human-defined masks; however, this approach is not suitable for automatic background removal in images of various flower types. IS-Net was easy to integrate but tended to produce unknown gray patterns in some removal results and also suffered from longer processing times (0.9-1.0s). BiRefNet excelled with high-quality background removal, faster processing times (0.1-0.2s), and straightforward integration, making it ideal for real-time applications. Based on these criteria, BiRefNet was chosen as our primary solution for background removal in generated images due to its superior overall performance and integration capabilities.

3.2.3 Summary

Throughout our exploration of interactive art content generation, significant emphasis has been placed on evaluating various image generation and post-processing models to determine their suitability for real-time, motion-driven art creation. Our evaluations of diffusion models, specifically the SDXL enhanced with ControlNet, have shown great promise in producing high-quality images rapidly. To further align with the interactive demands of our project, we incorporated the LCM-LoRA, which significantly reduced generation times. Additionally, BiRefNet was integrated for post-processing to effectively eliminate unwanted background elements, thereby improving the clarity and emphasis on the main subjects of the artwork.

This approach ensures both high quality and control over image generation, meeting the response times essential for interactive experiences. The development has resulted in a promising generation pipeline that holds potential to effectively utilize user motion to create engaging and visually appealing interactive art, particularly tailored to our focus on flower imagery. Based on these findings, we decided to move forward with this concept – interactive art content generation with motion, using the insights gathered as the foundation for further implementation and refinement.

4 MotionCanvas Prototype

In this chapter, we present the implementation details of a proof-of-concept MotionCanvas prototype.

4.1 System Overview

With the insights gained from the exploratory study, we present MotionCanvas which enables the translation of simple user movements into fine-styled artworks—specifically, flower paintings in our prototype—allowing users to have engaging real-time art creation experiences. Beyond accommodating individual user inputs, MotionCanvas also supports co-creation among multiple users within the same space. This functionality enhances the interactive experience and fosters community environment, further encouraging the collaborative engagement in artistic creation. Figure 4.1 illustrates how users interact with the system for a co-creation task of creating flowers.



Figure 4.1 Prototype of MotionCanvas (Co-creation scene): Users generate flower paintings through their movements, collaboratively creating a single artwork in real-time.

4.1.1 Hardware Settings

The hardware configuration of MotionCanvas is structured into three main components: the user interaction area, the local PCs, and the motion capture setup. The interaction area is equipped with a large screen measuring 165 cm × 92.8 cm as shown in Figure 4.2, which is connected to a local PC (PC1). This screen displays the interactive content and serves as the primary interface for user engagement. Additionally, three high-resolution cameras are strategically positioned to capture comprehensive video streams for motion detection.

The motion capture system operates on a PC (PC2), equipped with an Intel Core i9-14900KF processor and an NVIDIA GeForce GTX 4090 24 GB. Simultaneously, the generative model for creating flower paintings runs on PC1 with additional computing power from a remote server equipped with an NVIDIA GeForce GTX 4090 24 GB, facilitating enhanced performance and speed in art generation.



Figure 4.2 Hardware settings of the interaction screen and cameras.

4.1.2 Motion Capture System

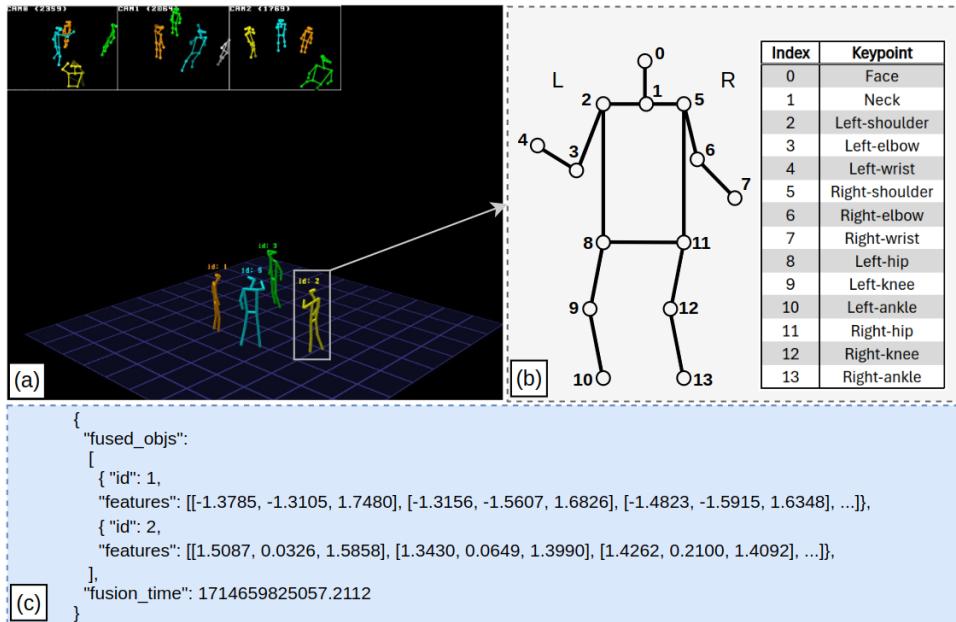


Figure 4.3 Motion Capture System: (a) Visualization result of 3D-pose reconstruction, (b) Keypoint details of the skeleton model, (c) JSON data including 3D coordinates (x, y, z) of 14 body keypoints and timestamps received from the system.

As mentioned in section 4.1.1, we employed a 3-camera installation for the motion capture system. The system utilizes a Transformer-based approach to reconstruct the 3D pose from the video feed in real time. Figure 4.3(a) shows the visualization result of the real-time reconstructed 3D-poses, displaying each user in the tracking area with a unique ID. Figure 4.3(b) illustrates the skeleton model, which consists of 14 body keypoints in 3D space. Figure 4.3(c) details the data format in JSON received from the system, representing the position of each individual as a 14×3 vector along with their ID and timestamp. The data refresh rate is approximately 15 frames per second.

4.1.3 Software System

Figure 4.4 shows the workflow of our MotionCanvas system. The system architecture integrates multiple components operating on both local and server-based platforms. The motion capture system is hosted on a PC running Ubuntu 22.04.4 LTS, while the Python motion processing program and the Unity program (developed using Unity 2022.3.19f1) are installed on another PC equipped with Ubuntu 20.04.6 LTS. The flower image generation pipeline is hosted on the server and executed within a Docker container, which includes essential libraries and frameworks such as Diffusers, Pytorch, and Flask¹, with Ubuntu 22.04.4 LTS and CUDA 12.3.

Local motion data communication between programs (Figure 4.4 ② to ⑤) is managed through an MQTT broker. Remote data exchanges between the local PC and the server(Figure 4.4 ⑥ to ⑧) occur via HTTP requests managed by a Python Flask API.

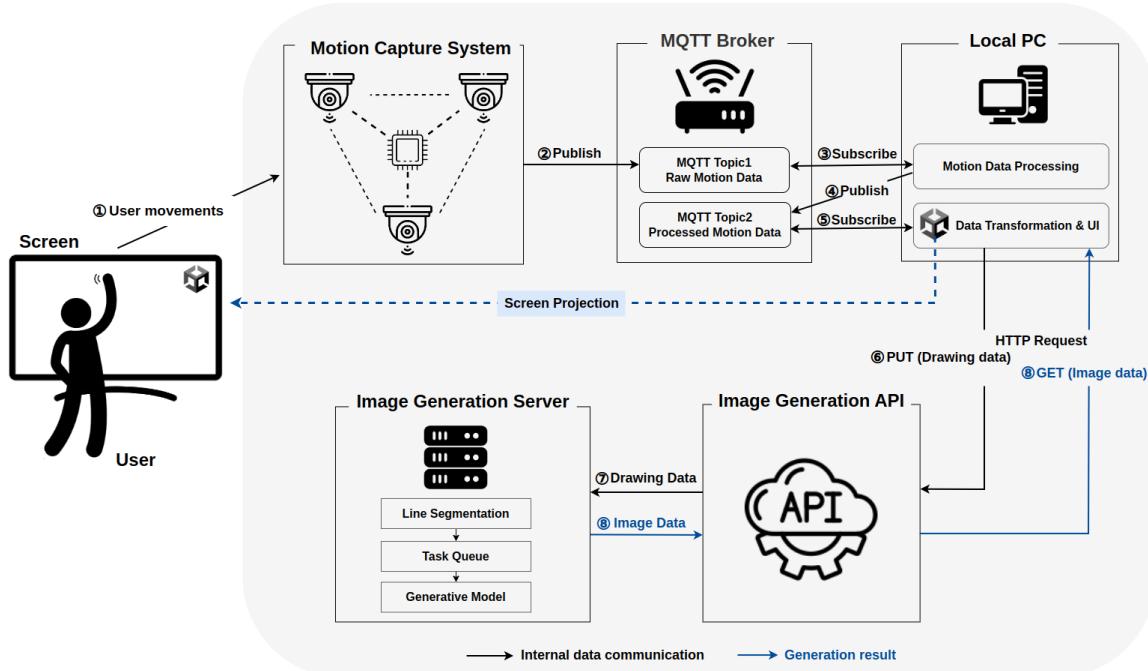


Figure 4.4 System Flow Overview.

The workflow comprises the following sequential steps:

1. Users interact with the system.
2. The motion capture system captures user movements and publishes the raw motion data to MQTT topic 1.
3. A Python program subscribes to topic 1, receiving the raw motion data.
4. This program processes the data, extracting relevant features, and publishes it to MQTT topic 2.
5. The Unity program, running on the same PC, subscribes to topic 2 and integrates the processed motion data.
6. Necessary calibrations and transformations convert 3D motion data into 2D drawing inputs within the Unity program. The transformed drawing data, including line points and positions, are sent to the Image Generation API via an HTTP PUT request.

¹ <https://flask.palletsprojects.com/>

7. The server-side flower image generation pipeline processes the drawing data, generating flower images.
8. The resulting images are returned to the local Unity program via HTTP GET request. Images are then displayed on the screen, rendered according to specified dimensions and positions.

4.2 Motion-to-Line Transformation Algorithm

In this project, we developed a Motion-to-Line Transformation Algorithm to enable users to draw in a 3D virtual space using motion capture data. The algorithm maps 3D motion data into a 2D space, tailored to the dimensions of the tracking area and the canvas. Figure 4.5 shows the two distinct directional views of a participant along with a corresponding 2D line drawn on the canvas.

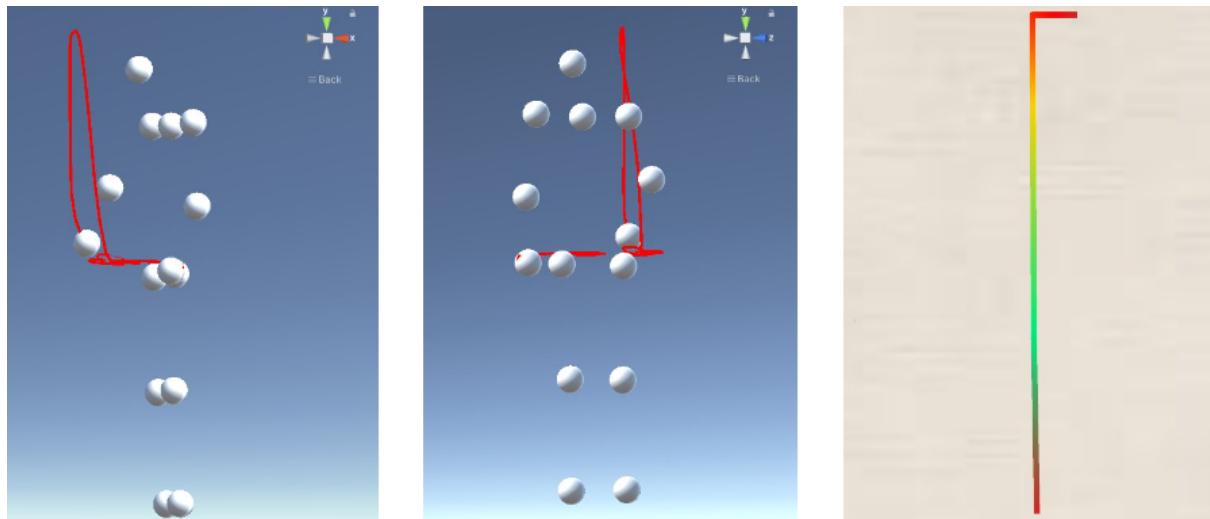


Figure 4.5 Motion to Line Transformation.

4.2.1 Data Collection and Processing

We receive a pose vector consisting of 14×3 scalar values from the motion capture system for each user. This data includes user IDs and joint positions, representing the movement of each user's body in the 3D space.

Using Unity, we create human models for each user based on their ID and we transform the joint positions from the motion capture system's coordinate system to Unity's coordinate system. Finally, we update the position of the corresponding joint Game Object in Unity's scene to reflect the new position calculated from the motion capture data.

By continuously updating the positions, we ensure that the human models accurately reflect the movements captured by the motion capture system in real-time.

4.2.2 Drawing Mechanism

We developed a Pen class in Unity for visualizing hand movements in 3D and transforming these movements into 2D representations. The Pen class employs 3D LineRenderers to accurately depict the trajectory of the user's hand movements, facilitated by real-time updates to the pen's position. To enhance reusability and ensure consistent configuration, the Pen class has been encapsulated as a Pen prefab. This prefab can be seamlessly attached to each hand of the virtual human models, maintaining uniformity in their setup.

Key components of the Pen class include:

1. **Initialization:** Methods for initializing the pen with specific user configurations.
2. **Line position updates:** Functions for updating the positions of the lines based on the pen's movement.
3. **Pen state management:** Mechanisms for managing the state of the pen, determining whether it is touching the canvas.

For the management of pen state, a threshold mechanism is implemented in Unity to determine when the pen makes contact with the drawing canvas. If the pen touches the canvas twice consecutively, the system activates a procedure to render the line in 2D while continuously updating the drawing canvas texture.

Furthermore, the system is designed to communicate with the Python server, enhancing its functionality by providing real-time data processing capabilities. The detailed drawing data sent includes:

- Start and end coordinates of the pen strokes,
- User IDs,
- Pen state.

This communication is efficiently managed through asynchronous UnityWebRequest calls, ensuring not only efficient data transfer but also real-time updates to the drawing task queue for subsequent processing in the flower image generation pipeline.

4.3 Flower Image Generation Pipeline

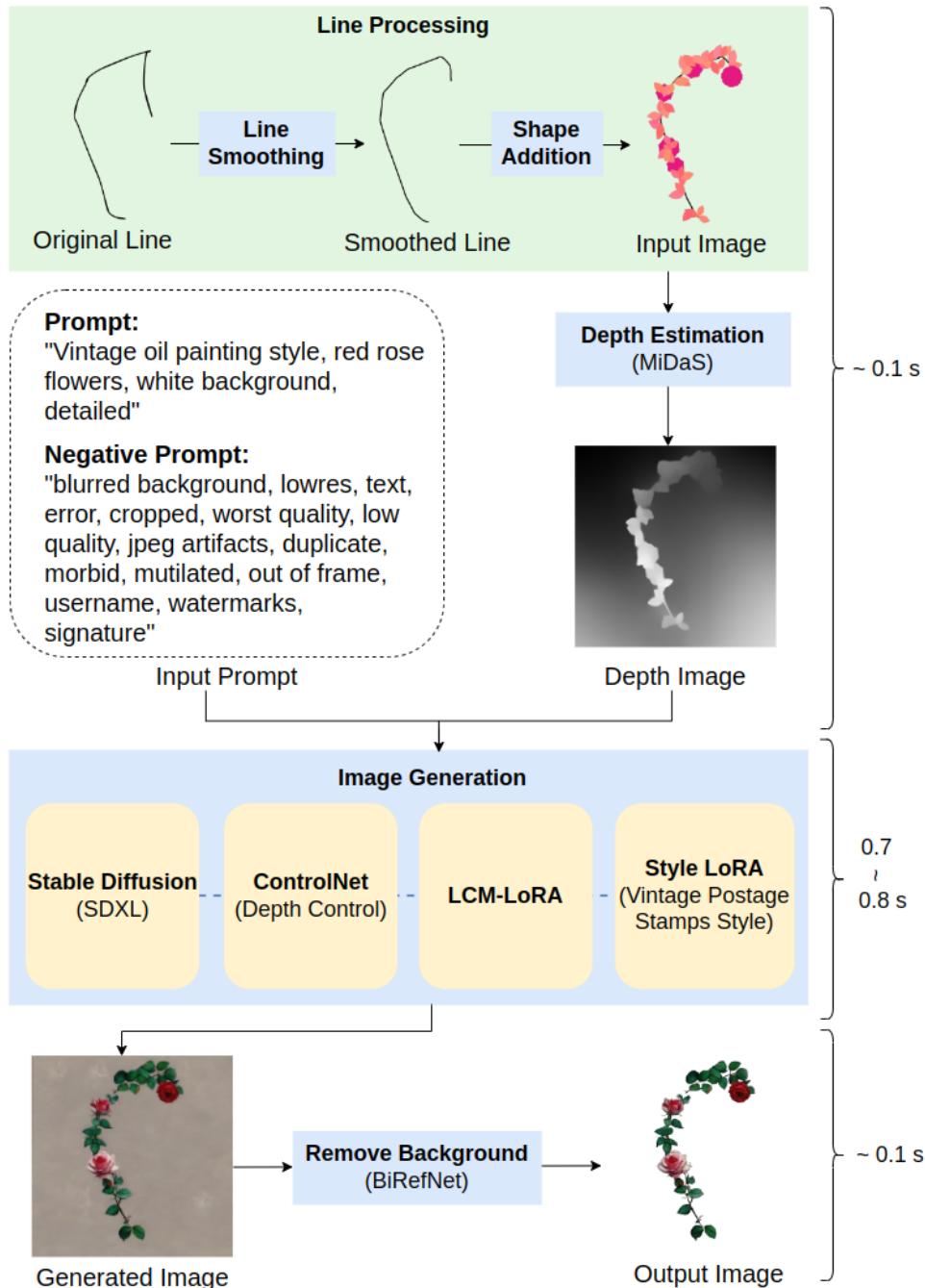


Figure 4.6 Flower Image Generation Pipeline

Figure 4.6 shows the flower image generation pipeline of MotionCanvas. This pipeline converts user-drawn lines into artistically refined flower images. We developed a line processing pipeline to create improved base input images for the generative model, as well as a stable diffusion-based pipeline for generating the final images. The whole pipeline takes about 1 second from receiving the input data to generating the final output image. Some examples of outputs from the pipeline are presented in Figure 4.7. The technical details of these pipelines are discussed in the subsequent sections.

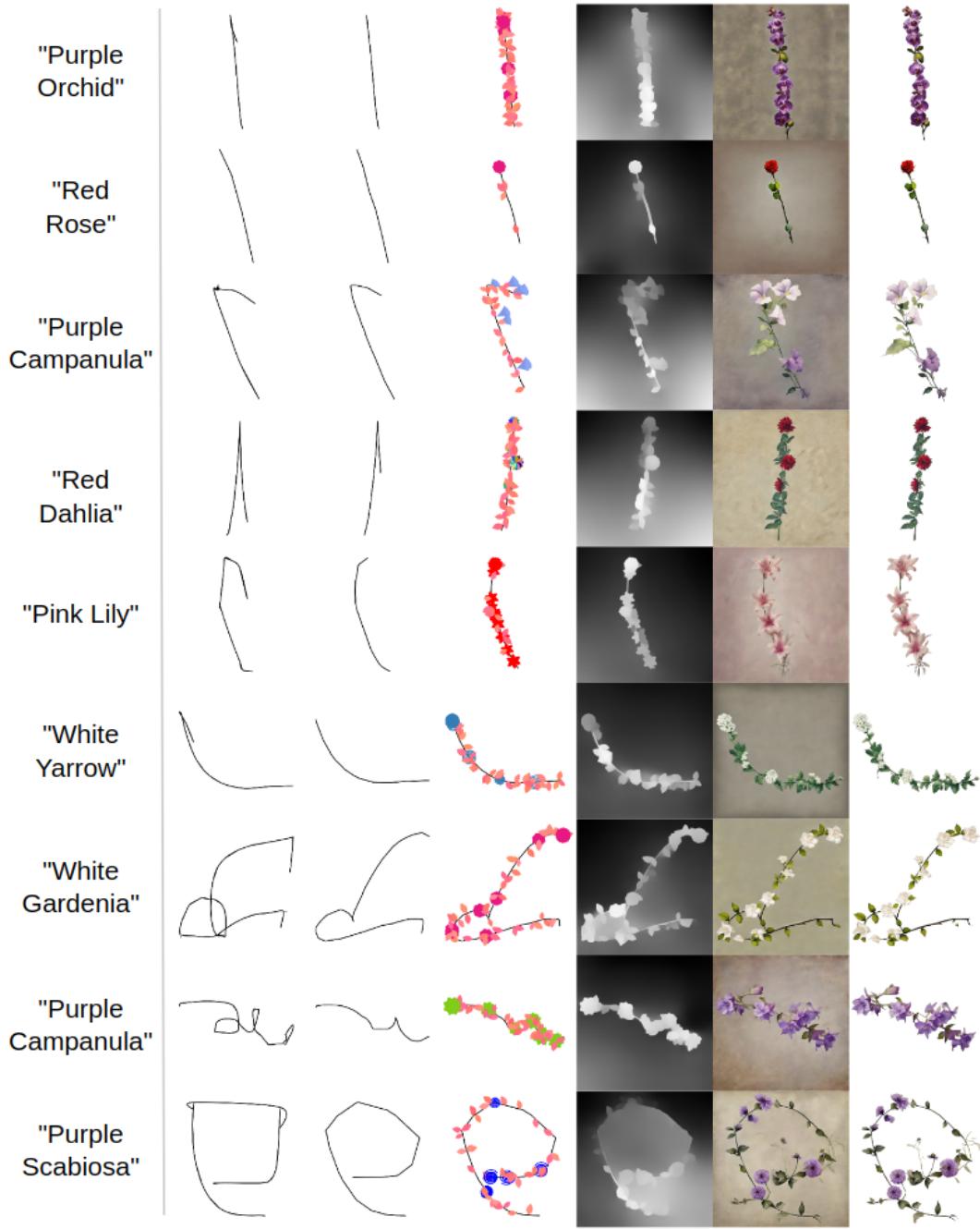


Figure 4.7 Example Outputs of Flower Image Generation Pipeline: From left to right, the columns represent: (1) the text prompt, (2) the original line drawing provided by the user, (3) the smoothed line, (4) the base line image with geometric shapes representing the flower, (5) the depth image input to the generative model, (6) the output image from the generative model, and (7) the final result image after applying the background removal function.

4.3.1 Line Processing

4.3.1.1 Line Smoothing Algorithm

The line smoothing algorithm includes three functions: *moving_average*, *filter_points* and *spline_interpolation*. These functions are designed to smooth and refine the input line, ensuring a clean and continuous trajectory suitable for further processing.

Moving_average: The function performs a smoothing operation on the input list of points that define the line as shown in Algorithm 1. The operation utilizes a moving average with a window size that dy-

namically adjusts based on the number of input points. For each point in the list, the function calculated the average of the x and y coordinates within the specified window. The averaged x and y values are then appended to a new list, which forms the smoother line. By applying moving average technique, the original line is smoothed by averaging values over the window, thereby reducing the sharp, unintended fluctuations.

Algorithm 1 Moving Average

```

1: function MOVINGAVERAGE(points, window_size=100)
2:   if length(points) < 6 then
3:     return points
4:   end if
5:   window_size ← min(round(length(points) × 0.2), window_size)
6:   smoothed_points ← []
7:   for each point i in points do
8:     avg_x ← average of x-coordinates in window [max(0, i – window_size + 1) : i + 1]
9:     avg_y ← average of y-coordinates in window [max(0, i – window_size + 1) : i + 1]
10:    append (avg_x, avg_y) to smoothed_points
11:   end for
12:   return smoothed_points
13: end function
  
```

Algorithm 2 Filter Points

```

1: function FILTERPOINTS(points, min_distance=3)
2:   filtered_points ← [points[0]]
3:   for each point i from 1 to length(points) - 1 do
4:     dist ← Euclidean distance between points[i] and filtered_points[-1]
5:     if dist > min_distance then
6:       append points[i] to filtered_points
7:     end if
8:   end for
9:   return filtered_points
10: end function
  
```

Filter_points: The function refines the smoothed line by ensuring that consecutive points are sufficiently spaced apart based on a pre-defined minimum distance between points as shown in Algorithm 2. This function eliminates closely-packed points, and improves the processing speed while preventing from the overly dense line segments.

Spline_interpolation: After having the evenly distributed line points with the *filter_points* function, to better form a smooth curve, we used an interpolation technique. As shown in Algorithm 3, the function takes four line points as input and calculates weighted averages to compute new intermediate control points, which is a concept used by cubic Bézier curve definitions [47]. To preserve the original shape of the line, we adopted the concept of Catmull-Rom splines [9], ensuring that the resulting curve passes through the start and end points of the original segment, as well as the newly calculated points.

4.3.1.2 Geometric Flower Shape Addition Algorithm

Learning from the exploratory study in section 3.2, we found that incorporating flower-like shapes in the image enhances controllability and improves generation results. Consequently, we developed a geometric flower shape addition algorithm to integrate geometric shapes representing flower petals and leaves into the line. This algorithm effectively transforms the abstract line into a preliminary floral structure.

Algorithm 3 Spline Interpolation

```
1: function SPLINEINTERPOLATION( $p_0, p_1, p_2, p_3$ )
2:    $c_1 \leftarrow p_1$ 
3:    $c_2 \leftarrow \frac{-p_0+6 \cdot p_1+p_2}{6}$ 
4:    $c_3 \leftarrow \frac{p_1+6 \cdot p_2-p_3}{6}$ 
5:    $c_4 \leftarrow p_2$ 
6:   return  $c_1, c_2, c_3, c_4$ 
7: end function
```

Geometric Shape Creation

The geometric shapes and art are often closely related. Drawing from this insight, we conducted a survey on common flowers and subsequently designed seven geometric shapes to represent different petals (Figure 4.8 (a)-(g)), along with one shape to represent a flower leaf (Figure 4.8 (h)) for the prototype.

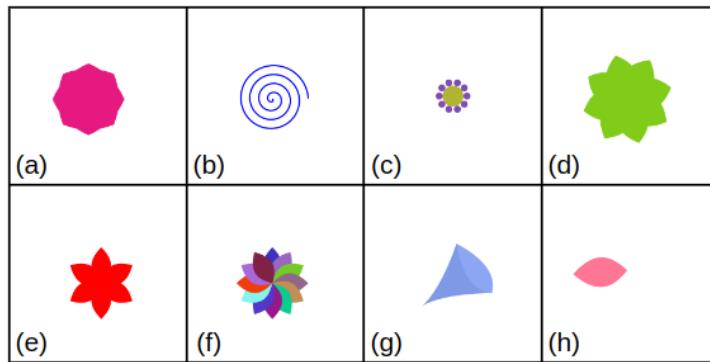


Figure 4.8 Design of Geometric Shapes for Different Flowers.

Drawing Method

We implemented eight distinct functions, each responsible for drawing one of the geometric shapes along the line. These shapes are strategically placed to achieve a specific visual effect that mimics the organic and varied nature of floral arrangements. Shapes are added with variations in size, rotation, and density. These variations are randomly selected within predefined ranges to create naturally occurring patterns. This randomness ensures that no two arrangements are exactly alike, similar to how no two flowers in nature are identical. The placement allows for overlaps and varying orientations, enhancing the realism and aesthetic quality of the generated floral images.

All drawing functions are built using the Pycairo² library, which provides the good support for creating vector-based complex graphics from scratch with various drawing commands. The vector-based nature of Pycairo ensures that it can produce images with smooth lines and consistent quality, regardless of the resolution. This makes it an excellent choice for projects that require precise and scalable graphics.

Geometric Shape - Flower Pairing

To facilitate the pairing of geometric shapes with corresponding flower types, we created a JSON file. This file maps each drawing function to specific flower types, as shown in Table 4.1. The program reads this JSON file and randomly selects a pair of drawing function and flower type. This selected flower type is then used as part of the text prompt for the image generation process.

² <https://pypi.org/project/pycairo/>

Table 4.1 Patterns and Associated Flowers

| Pattern | ID | Function Names | Flowers |
|---------|----|--|--|
| ✿ | 1 | draw_overlapping_petals | Pink Rose, Red Rose, Pink Peony, Red Peony, Red Begonia, White Gardenia, Red Petunia, Purple Petunia, Purple Orchid |
| 螺旋 | 2 | draw_spiral, draw_nested_spirals | White Yarrow, Purple Scabiosa, White Dandelion |
| ✿ | 3 | draw_flower_with_central_pattern | White Daisy, Yellow Daisy, Yellow Sunflower, White Aster, Purple Aster, Purple Orchid, Purple Cosmos, Yellow Calendula |
| ✿, * | 4 | draw_rotated_layers_flower, draw_simple_flower | White Lily, Pink Lily, Yellow Daffodil, White Freesia, Purple Campanula, Red Amaryllis |
| ✿ | 5 | draw_radial_symmetry_flower | Red Dahlia, Yellow Dahlia, Pink Dahlia, Yellow Chrysanthemum, Pink, Yellow Carnation, Purple Cornflower, White Jasmine |
| ◁ | 6 | draw_neiles_parabola | White Lily, Pink Lily, Purple Crocus, White Crocus, Yellow Crocus, Purple Morning Glory |

Figure 4.9 presents examples of the output results after applying the complete line processing program.

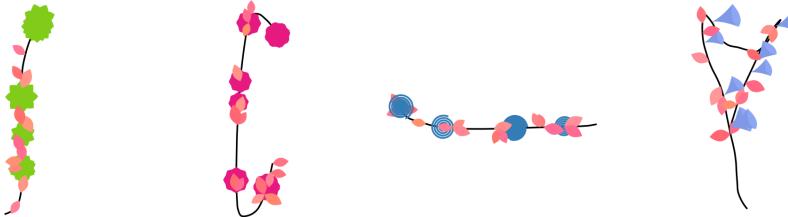


Figure 4.9 Examples of Line Processing Results.

4.3.2 Stable Diffusion Model Configuration

Building on the insights from the exploration study as mentioned in section 3.2.3, we have enhanced our image generation pipeline by integrating several components: SDXL, ControlNet, LCM-LoRA, and a style LoRA module. SDXL serves as the foundational text-to-image model. ControlNet extends this by incorporating depth images as additional conditioning inputs, enhancing the contextual relevance of generated images. LCM-LoRA accelerates the image generation process, improving efficiency. For the style-specific outputs, we utilize a style LoRA model, which is adept at producing images in a vintage art style, thereby aligning perfectly with our project requirements.

4.3.2.1 Model Construction Configuration

All modules within the pipeline have been finely tuned to optimize performance and output quality for the project. In technical details, we utilized the **StableDiffusionXLControlNetPipeline** from the Diffusers library to construct the SDXL model. This pipeline comprises four integral components, detailed in Table 4.2, with explicit configurations for each. Additionally, we incorporated a merging of the two LoRA weights: the LCM-LoRA and the style-specific LoRA. The detailed configurations for this merging process are presented in Table 4.3.

Table 4.2 StableDiffusionXLControlNetPipeline Configuration

| Parameter | Value | Description |
|-------------|--|--|
| vae | madebyollin/sdxl-vae-fp16-fix | Variational autoencoder for image reconstruction |
| sd_model | stabilityai/stable-diffusion-xl-base-1.0 | Core text-to-image generation model |
| control_net | diffusers/controlnet-depth-sdxl-1.0 | Depth-based conditioning input model |
| torch_dtype | torch.float16 | Precision type for model computations |

Table 4.3 LoRA Configuration

| Parameter | Value | Description |
|------------------|------------------------------------|---|
| lora | KappaNeuro/vintage-postage-stamps | Primary LoRA model for vintage style |
| lora_weight | Vintage Postage Stamps.safetensors | Weight file for primary LoRA model |
| adapter_name | vintage | Adapter name for primary LoRA model |
| adapter_weight | 0.9 | Blending weight for primary LoRA model |
| lora_2 | latent-consistency/lcm-lora-sdxl | Secondary LoRA model for model acceleration |
| lora_weight_2 | pytorch_lora_weights.safetensors | Weight file for secondary LoRA model |
| adapter_name_2 | lcm | Adapter name for secondary LoRA model |
| adapter_weight_2 | 1.0 | Blending weight for secondary LoRA model |
| scheduler | LCMScheduler | Scheduler for managing inference steps and blending |

4.3.2.2 Model Inference Configuration

The image generation model utilizes text prompts and a conditioning depth image as inputs, along with other inference parameters such as inference steps and conditioning scale. The text prompts consist of a positive prompt, which guides the generation content, and a negative prompt, which specifies undesirable content to be avoided in the generated result. In our project, the positive text prompt and the conditioning image vary according to the input data. Conversely, the negative prompt and the remaining inference parameters are fine-tuned for the project and remain constant throughout the process. The content of the negative prompt is shown in Figure 4.6. Details of the remaining configurations are provided in Table 4.4.

Table 4.4 Inference Parameters Configuration

| Parameter | Value | Description |
|-------------------------------|-------|---|
| num_images_per_prompt | 1 | Number of images generated per prompt |
| num_inference_steps | 4 | Number of steps during inference |
| guidance_scale | 1 | Scale factor for guidance during generation |
| controlnet_conditioning_scale | 0.9 | Scale for conditioning with ControlNet |
| control_guidance_end | 0.7 | Endpoint for control guidance |
| lora_scale | 0.9 | Scale factor for LoRA |

4.3.3 Task Queue Management

As previously mentioned, each generation task takes approximately one second to complete. When users continue drawing, new tasks are continuously initiated. To ensure tasks processed sequentially and to prevent the system overload, we implemented a *TaskExecutor* class. This class uses queuing mechanisms and threading to manage task execution efficiently.

The *TaskExecutor* initializes a task queue and starts a worker thread, which runs in the background and stops when the program exits. All incoming tasks are added to the queue first. The worker thread continuously retrieves and processes tasks from the queue using the pipeline shown in Figure 4.6. If task processing returns "*Invalid*", a default image is set; otherwise, the processed image is stored and sent to the Unity program upon HTTP requests later.

4.4 User Interface Design

A well-designed user interface enhances user engagement and improves the overall experience. Given our project's theme of allowing users to create their own flower art through motion, we selected an elaborate painting frame with a beige-colored fabric texture canvas as shown in Figure 4.10(a). This choice aims to evoke an art gallery ambiance and stimulate creativity. Additionally, instead of plain black lines, we utilized rainbow-colored lines by applying the gradient variable of LineRenderer in Unity as shown in Figure 4.10(b). Inspired by common website loading animations, we implemented an animation effect that loops the gradient around the line, thereby reducing users' perceived latency.



Figure 4.10 User Interface Design: (a) Drawing canvas with painting frame, (b) Rainbow-colored lines.

5 User Evaluation

User testing is a critical part of the development process for multi-user interaction systems, ensuring their usability, functionality, and effectiveness in collaborative experiences. This chapter outlines the user testing methodology and shows valuable feedback from users to iterate and optimize the system's design.

5.1 Setup

The experiment was conducted in a controlled interaction area which was demarcated in front of the display screen. The area was approximately 1.5 meters in length and 1.2 meters in width. Figure 5.1 shows the test apparatus and environment.

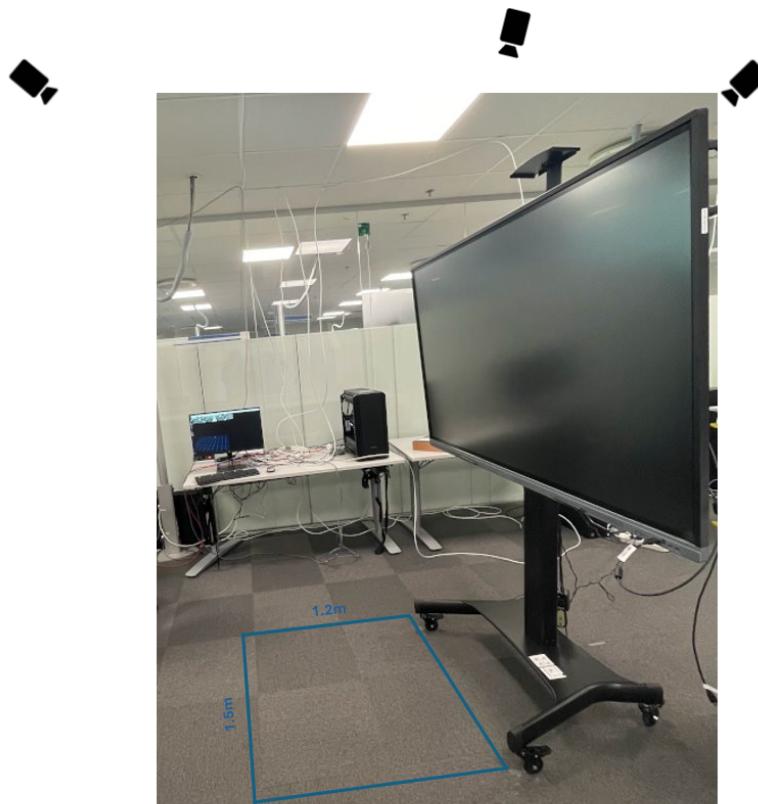


Figure 5.1 User Test Environment.

This area provided participants with sufficient space to move and interact comfortably while ensuring optimal visibility of the interactive content on the screen. The size and layout of the interaction area were chosen to facilitate natural and intuitive user interactions within the constraints of the test environment.

5.2 Participants

21 participants (14 male, 7 female) took part in the user test and we received 20 responses. The participants' ages varied between 21 and 34 years old. They were from VR/AR, computer science, physical geography and ecosystem science, linguistic, and some were software engineers, and machine learning engineers.

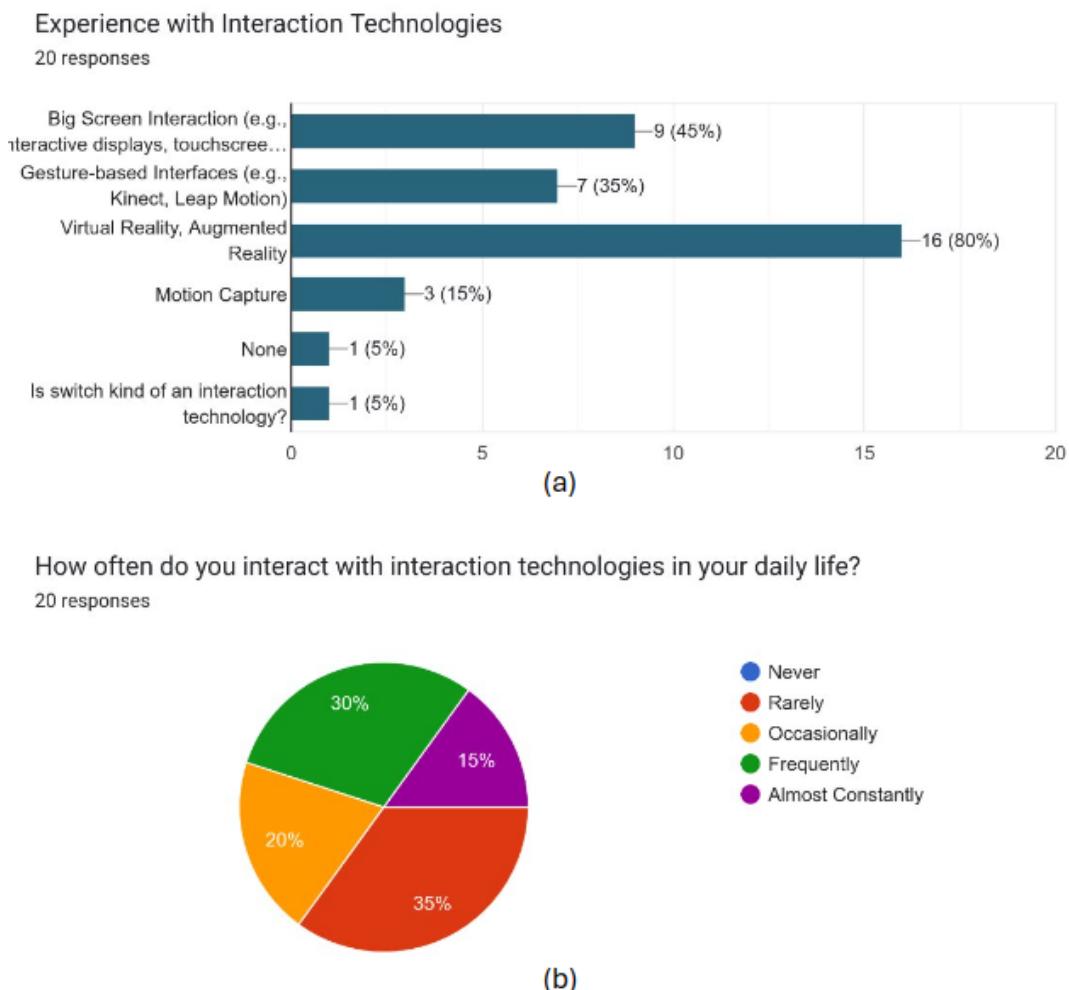


Figure 5.2 Experience with Interaction Technologies.

As shown in Figure 5.2 and 5.3, most participants used interactive product, such as big screen interaction, VR, AR and gesture-based interaction. Most of them had experienced AI interaction and around half of them had interacted frequently.

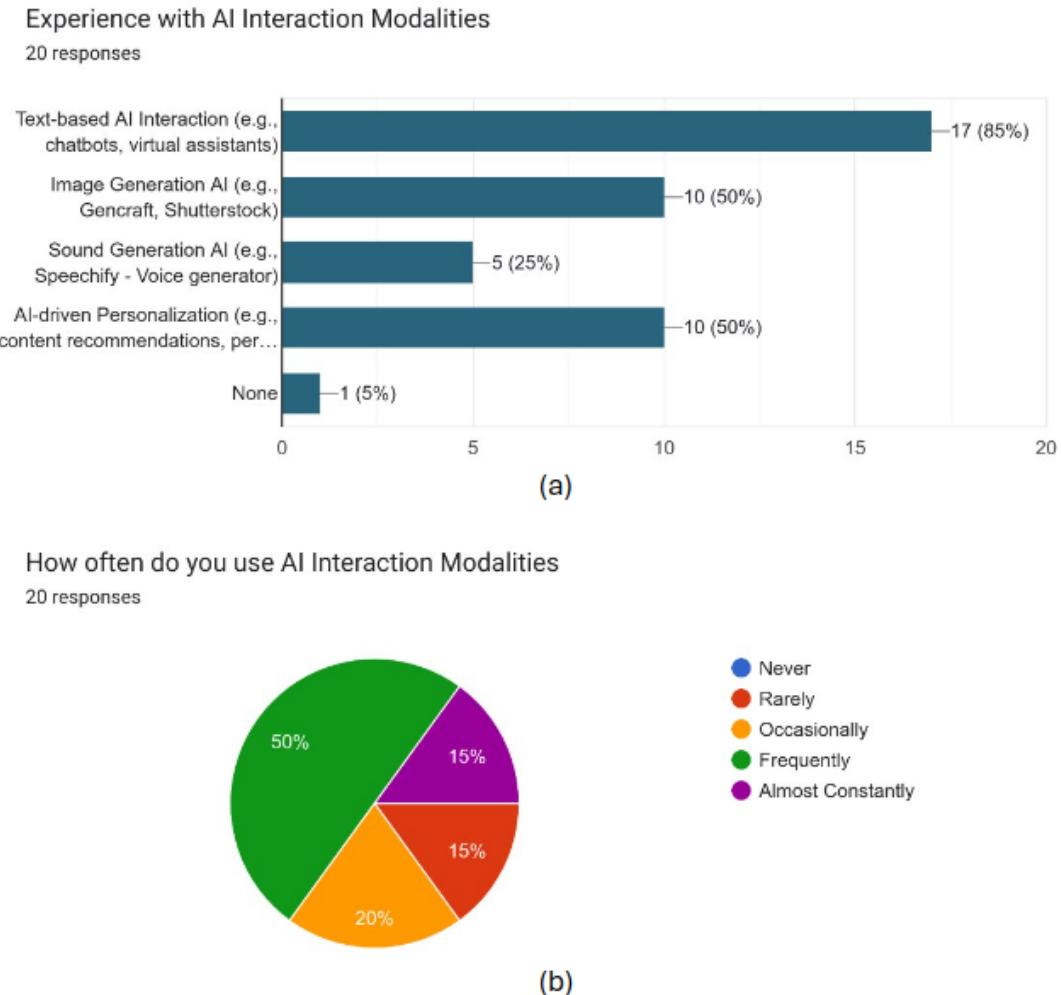


Figure 5.3 Experience with AI Interaction Modalities.

5.3 Methods

We conducted a preliminary experiment to test our project. From our functions and use cases, we focus on three tasks to evaluate our project: exploration task, standardized task and co-creation task. These tasks were designed to assess various functions and use cases of the system. Participants experienced two single-person tasks and one multi-person task.

As an initial evaluation, we used questionnaires to gather feedback of each task and SUS as well as subjective preferences. The questionnaires for each task included quantity questions on ease, expression and enjoyable and quality questions on challenges and difficulties. This approach allowed us to capture both numerical data and in-depth insights into the user experience.

5.4 Procedure and Tasks

Prior to the usability testing, participants were explicitly informed that their data would be utilized exclusively for research purposes. A consent form was presented, detailing the objectives of the study, the nature of their involvement, and the manner in which their data would be utilized. To facilitate a comprehensive understanding of the drawing function and inspire creativity, participants were provided with an example of drawing animations before commencing the tasks. Following this, participants were presented with the task sheets outlining the specific activities they would be required to complete.

The user testing comprised three distinct tasks, which were designed to assess the usability and functionality of the system. Upon completion of the tasks, participants were asked to complete two questionnaires: one aimed at gathering feedback and the other assessing the SUS scores of the system. The full questionnaires are presented in Tables A.2 and A.3 in the Appendix.

5.4.1 Exploration Task

We first asked participants to create their own drawings freely to explore different hand motions and interactions within the defined interaction area, as shown in Figure 5.4. They were encouraged to experiment with various movements using their fingers, hands and fists to express their creativity. Additionally, participants were encouraged to use both hands simultaneously to interact, allowing for more dynamic and expressive interactions. By exploring the system's capabilities and limitations, participants gain a deeper understanding of how to interact effectively while uncovering any potential challenges or difficulties. This task aimed to familiarize users with the system and to measure how well to express creativity through various interactions.



Figure 5.4 Exploration Task.

As participants engage in creative exploration, we observed how well the system accommodates a range of movements. This includes assessing the system's responsiveness to different hand motions and its ability to interpret user intentions accurately. Additionally, we paid close attention to any difficulties or frustrations encountered by participants, providing insights into areas that may require improvement. Figure 5.5 shows some examples of the participants' creations.



Figure 5.5 Some Results of Exploration Task.

5.4.2 Standardized Task

The Standardized Task presented participants with a specific challenge: to draw pre-defined patterns as shown in Figure 5.6. The predetermined patterns include vertical lines, poly lines, curves and circles. This task evaluated the accuracy of the motion capture system in interpreting user gestures, as well as participants' ability to perform structured tasks within the interactive environment.

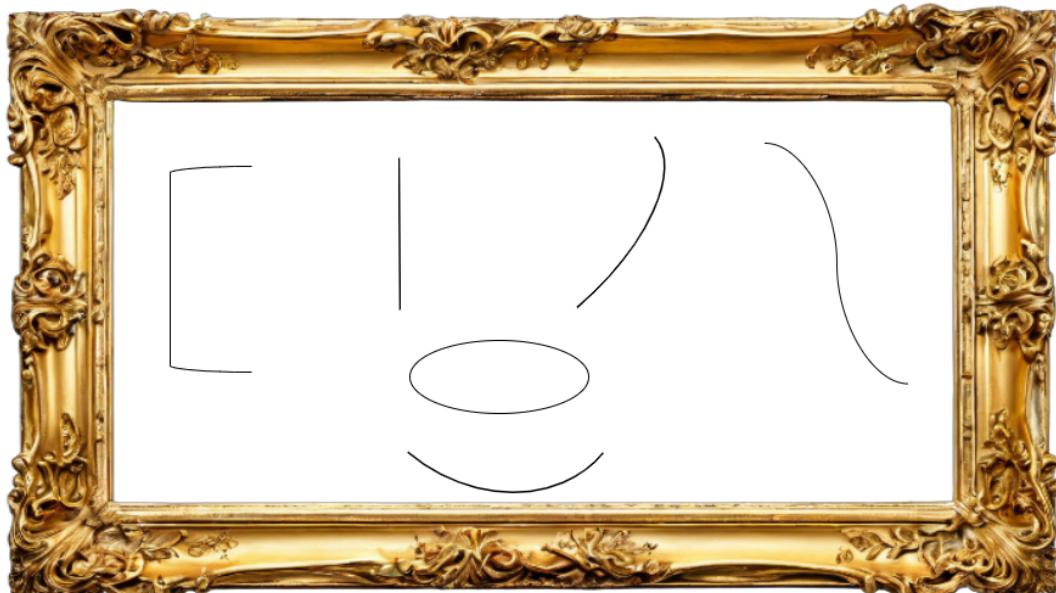


Figure 5.6 Standardized Task.

Participants were asked to pay attention to the ease of drawing each pre-defined pattern, taking into consideration factors such as the responsiveness of the system to their hand movements, any offset or delay between their gestures and the corresponding actions on the screen, and the overall difficulty level of drawing the patterns accurately. Feedback and suggestions were gathered to improve the responsiveness and accuracy of the system.

Additionally, participants were encouraged to provide feedback on quality of generated content, any challenges and limitations experienced, and suggestions for improvement. By gathering these feedback,

we were able to iterate and improve the interaction system. Figure 5.7 shows some examples of the participants' creations.



Figure 5.7 Some Results of Standardized Task.

5.4.3 Co-Creation Task

In the Co-Creation Task, participants worked with a partner to draw a set of pre-defined patterns, as shown in Figure 5.8. Due to the limitation of screen size, two participants worked together. These patterns served as inspiration and guidance for participants and they were allowed for creative modification. Participants were encouraged to work together, simultaneously creating their own lines and shapes, and combining them to form a cohesive design.



Figure 5.8 Co-creation Task.

The aim of this task was to foster collaboration and communication between participants while leveraging their individual creativity to produce a unique pattern. Participants were encouraged to discuss ideas, make a plan, and coordinate with partners. Additionally, participants were encouraged to iterate and refine their creation with the freedom to add more lines or redraw and cover existing lines if the

initial result did not meet their satisfaction. This iterative approach encouraged experimentation and exploration, empowering participants to take ownership of the creative process and make adjustments based on their preferences and insights gained during collaboration.

Participants provided feedback on the effectiveness of the collaborative interaction, the quality of generated content, which way (single-person or multi-person) participants prefer and the part they enjoyed and struggled. Figure 5.9 shows some examples of participants' creations.



Figure 5.9 Some Results of Co-creation Task.

5.5 Results

20 participants provided their feedback for each task by quantity and quality questions and open-ended questions. And we used SUS to evaluate to whole system. Participants' SUS scores were used to analyzed the system's overall usability and identify areas for improvement.

The expectations from participants for a project that combines human motion with GenAI to create artistic works included creative and interesting content, quick and accurate responses, easy to use and they showed curiosity and excitement of the project.

Participants specifically noted:

- "Could generate interesting and unique results" and "create cool artistic things" as their expectations.
- They are "curious about how good the tools are."
- The performance of "easy to use, visually artistic and fun experience" and "quick and accurate".

5.5.1 Exploration Task Feedback

For the exploration task, we gathered the feedback on several aspects including ease of understand and interact with system, impact of motions on the generated content, satisfaction with using hand motions to create flowers and enjoyment of exploration task. Additionally, participants were asked an open-ended question on challenges and difficulties.

Figure 5.10 shows that half participants considered it was easy to understand and interact with the system. A significant number of participants expressed difficulty, the main reason was shortcomings in the introduction of our project and the guidance video provided.

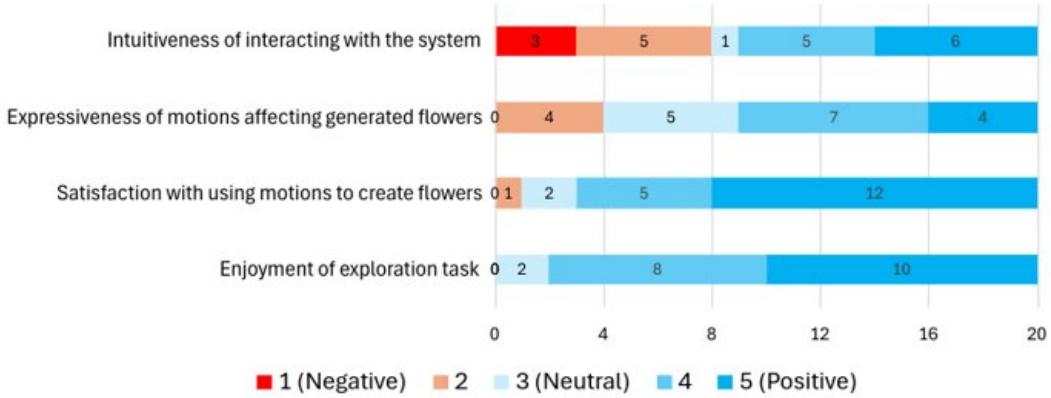


Figure 5.10 Exploration Task Evaluation.

Most participants considered that hand movements had influence on the resulting flowers. This suggests that participants felt some impact from their hand motions but not all the time.

Satisfaction with using hand to create flowers levels are high, with the majority of participants (12) giving the highest scale. This indicates that most participants found using hand to create flowers very enjoyable. Only one person did not enjoy it much.

Enjoyment of the exploration task is generally high. The majority of participants rated their satisfaction at four or five and two people were less satisfied (ratings three).

Participants had some challenges and difficulties interacting with the system, such as hard to start and end the lines, annoying offset, limited interaction area and latency of drawing.

5.5.2 Standardized Task Feedback

After completing the exploration task, participants became familiar with the system and they were asked to draw a pre-defined pattern. In this standardized task, we focused on the ease and accuracy of line drawing, appeal of the generated flowers, satisfaction with response time, fit of the generated flowers with hand movements, and enjoyment of the standardized task.

Figure 5.11 shows that most participants found the ease and accuracy of line drawing certainly challenging, with a majority rating it between two and four. The highest concentration is at rating three, and the rating of this task is more concentrated compared with last task. One person number found it very easy (rating one) and one person found it very hard (rating five).

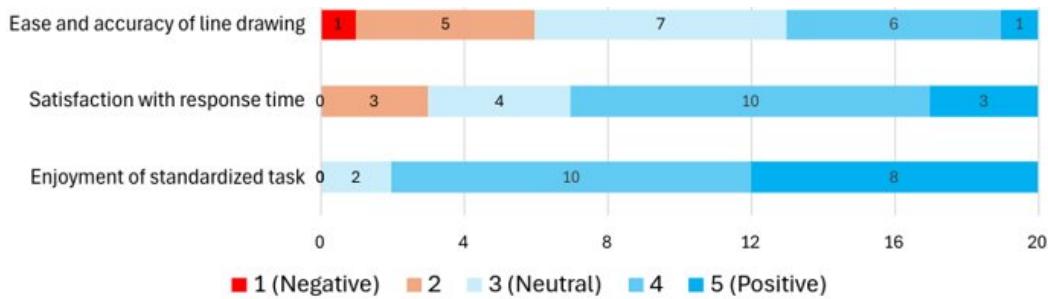


Figure 5.11 Standardized Task Evaluation.

Satisfaction with response time is high, with 13 participants rated it a four or five. There were still three participants rating two and four participants rating three.

Participants generally enjoyed the standardized task, the majority rating it a four or five. There were a few lower ratings.

In this task, participants met some challenges, such as hard to draw the circle, wrong touching by another hand, hard to start and end lines, latency of drawing and offset between lines and hands. There were two types of latency: the latency between motion and drawing, and the latency between drawing and generated flower paintings. Users found the first type of latency—between their motion and the resulting drawing on screen—noticeably challenging. The main reason is the latency from the motion capture system. Additionally, using the big screen means that we need re-calibration after every screen movement, making it difficult to eliminate offset issues.

Participants provided some suggestions to improve the system. One common recommendation was the addition of an on-screen shadow effect or cursor to show where the system detects the user's hand position. This feature would help users better perceive and understand their hand movements, thereby improving overall precision and ease of use. Additionally, participants suggested adding an undo feature that allows them to correct mistakes by reversing their most recent drawing, enhancing flexibility and user control. Reducing latency and offset was identified as a critical area for improvement. Participants noted that minimizing delays and inaccuracies in the system's response to hand movements would significantly enhance the interaction experience. Furthermore, the addition of animations and sound feedback when the generated flowers appear, would make the experience more engaging and enjoyable.

Participants specifically noted:

- Add "a chance to undo a line if we made a mistake".
- The visual feedback by "tips on the screen."
- The sensory feedback "animations on flowers" and "add a cool sound".
- "Reduce latency and fixed the issue where the finger and the result are not in the same position"

In the part of generating flowers, we added smoothing and filtering algorithms, so we collected participants' experience feedback in this aspect. 13 participants said they felt this part of the change, and the rest said they were not sure. At the same time, one participant thought that it would also be interesting without smoothing and filtering. And we gather the feedback on generated flowers in Figure 5.12.



Figure 5.12 Generated Flowers Evaluation.

The appeal of the generated flowers is high, with most participants rating it a 4 or 5. This indicates that participants satisfied with flowers generated by GenAI. No users rated it at the lowest levels (1 or 2), suggesting overall positive feedback in this aspect.

Most users felt that the generated flowers fit very well with their hand movements, with the highest ratings (4 and 5) being predominant. This indicates a strong correlation between user hand movements and the resulting flowers, enhancing the interactive experience.

Generally, participants were satisfied with the generated flowers. The majority of participants rated their satisfaction at 4 or 5 and some people were less satisfied.

5.5.3 Co-creation Task Feedback

After Exploration and Standardized task, participants were asked to pair up with a partner to engage in a co-creation task. This task aimed to evaluate the ease of co-creation, enjoyment of co-creation and interest in future interaction. Additionally, open-ended questions were posed to gather qualitative insights into interesting aspects and participants' descriptions of the interactive system.

Figure 5.13 shows that nearly half of the participants (nine) found easy to co-create, a significant portion encountered difficulties, highlighting the need for improvements in the system's operability.

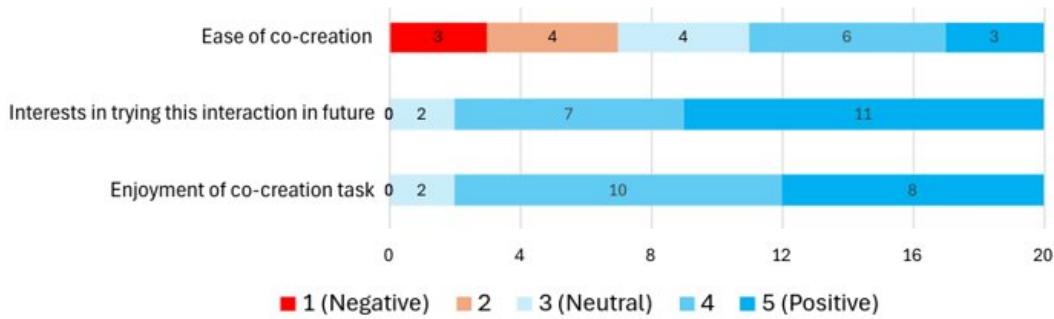


Figure 5.13 Co-creation Task Evaluation.

Interest in future interactions with the system was very high. 18 participants expressed being interested in engaging with the system again. Only two participants rated their interest as medium, indicating a generally positive outlook towards future use and the potential for sustained engagement with the system.

Participants' enjoyment of the co-creation task was high. 18 participants expressed positive sentiments about the collaborative aspect, indicating that the task was engaging and well-received. Only two participants rated their enjoyment as medium, suggesting further enhancements could ensure a consistently engaging experience for all users. Additionally, 11 participants expressed they enjoyed more in co-creation and 4 preferred single-person creation.

Participants were asked open-ended questions about the interesting aspects and their descriptions of the interactive system. Many participants expressed that it was very fun and interesting to see flowers emerging from their hand movements or drawings. They enjoyed the novelty and creativity of the interaction, finding it fascinating to watch their sketches transform into digital flower images. This visual feedback was particularly engaging, making the process feel magical and rewarding.

Participants also highlighted the challenge of drawing accurately with the system. Initially, many found it difficult to control their hand movements precisely, but after several attempts, they felt they began to "find the feeling" of how to interact with the system effectively. This learning curve, while initially frustrating for some, ultimately contributed to a sense of achievement as they mastered the interaction.

In describing the interactive system, participants emphasized the enjoyment of drawing pictures with friends and observing flowers generated along the lines they drew. The collaboration task was a significant highlight, with participants valuing the opportunity to communicate and co-create with a partner. This shared creative process allowed each participant to leverage their individual creativity effectively. Additionally, participants enjoyed the process of spontaneously assigning specific lines to themselves and discussing the perfection of particular lines and potential improvements.

Participants specifically noted:

- "Drawing the picture with my friends" is a favorite part of the experience.
- The visual delight of "seeing how flowers are generated on the place of the line I drew."
- The importance of "communication between users" during the co-creation process.
- "When I saw the flowers coming from my hand, I felt very happy."
- The enhanced drawing experience due to "the co-creation with a partner," and "share the creativity each one has." which made the activity more enjoyable and creatively stimulating.

Overall, the co-creation drawing experience, combined with the dynamic visual feedback, was seen as a key strength of the system, enhancing both enjoyment and creative expression. These insights will help future improvements to further refine the user experience and address the initial challenges faced by participants.

5.5.4 System Usability Scale

Based on the SUS responses from 20 participants for our system, we assess the usability of our project to show the user-friendliness and overall quality. Table 5.1 shows the results of the SUS.

Table 5.1 SUS Responses

| | I think that I would like to use this system frequently. | I found the system unnecessarily complex. | I thought the system was easy to use. | I think that I would need the support of a technical person to be able to use this system. | I found the various functions in this system were well integrated. | I thought there was too much inconsistency in this system. | I would imagine that most people would learn to use this system very quickly. | I found the system very cumbersome to use. | I felt very confident using the system. | I needed to learn a lot of things before I could get going with this system. |
|-------------------|--|---|---------------------------------------|--|--|--|---|--|---|--|
| Strongly Disagree | 0 | 7 | 0 | 10 | 0 | 5 | 0 | 8 | 0 | 13 |
| Disagree | 1 | 12 | 2 | 4 | 1 | 7 | 0 | 5 | 2 | 7 |
| Neutral | 8 | 1 | 1 | 4 | 5 | 3 | 3 | 6 | 2 | 0 |
| Agree | 7 | 0 | 11 | 0 | 8 | 4 | 4 | 1 | 7 | 0 |
| Strongly Agree | 4 | 0 | 6 | 2 | 6 | 1 | 13 | 0 | 8 | 0 |

* Numbers: the count of people giving this rating.

The SUS provides a single score on a scale from 0 to 100. A score of 68 is generally considered to be the average benchmark for usability. Scores above 68 indicate above-average usability, while scores below 68 suggest that there may be usability issues that need to be addressed. Our project's mean SUS score is 77 out of 100, indicating that our system has good usability. Users generally find it usable and tend to use it frequently. As shown in Figure 5.14, the SD = 15.84 with a minimum score of 37.5 and a maximum score of 100.

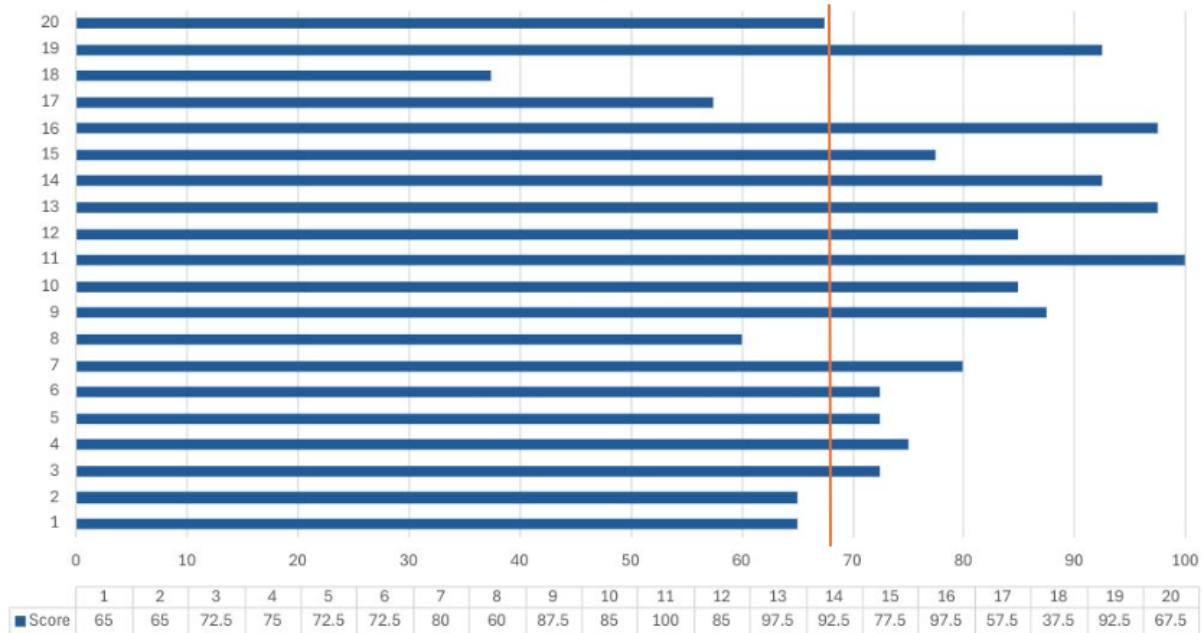


Figure 5.14 Co-creation Task SUS Evaluation.

Participants generally agreed that they would like to use the system frequently, indicating a positive attitude towards its usage. The majority of participants found the system easy to use and felt confident while using it. Most participants believed that others would learn to use the system quickly, suggesting that it has a low learning curve. The integration of various functions within the system was perceived positively, indicating coherence and efficiency.

While the overall usability is good, there are still areas for improvement. For example, some participants found the system unnecessarily complex, indicating a need for simplification or clearer instructions. A minority of participants felt that they would need the support of a technical person to use the system, suggesting that aspects of the system are not intuitive. There were also some concerns about inconsistency within the system, which could lead to confusion or frustration among participants. A small number of participants found the system cumbersome to use, indicating that there are aspects of the user interface or workflow that could be optimized for efficiency.

6 Discussion

In this chapter, we discuss the research questions based on the insights gained throughout the development process and the results from the user evaluation. We also outline the study's limitations and propose directions for future work.

6.1 Answer Research Question

In our exploration of integrating GenAI and motion capture technologies to create artistic interactive experiences, we addressed the research questions raised in Chapter 1. This section delves into our approach and findings for each question, highlighting the methods used and the results achieved.

Research Question 1: How to unify GenAI models with motion capture to create a new form of interactive art experience?

We capture human motion data as a pose vector consisting of 14 body joints within a 3D coordinate system. This ensures detailed and accurate representation of human movements. Hand movement data is extracted and sent to a server. This data is then converted into line drawings that follow six predefined patterns representing over 20 different types of flowers and leaves, commonly found in nature. The line patterns, along with depth information, are used as inputs for Stable Diffusion models. These models generate flower images based on the patterns and depth cues, creating interactive and dynamic visual content.

Research Question 2: How to ensure that the generated content aligns with the overall style and visual aesthetic of an interactive experience?

We explored various models and decided to use SD v1.5 and SDXL which are high-quality generative models, known for their exceptional image generation capabilities at resolutions of 512×512 and 1024×1024 , respectively. By integrating ControlNet, we can condition the generative models with additional inputs such as line scribbles, depth information, and human poses. This integration ensures that the generated content is coherent with the visual style dictated by user interactions and predefined artistic guidelines. Furthermore, Stable Diffusion uses depth information to generate flowers, which ensures stability and control over the generated content, resulting in reasonably shaped flowers with appropriate colors and sufficient leaves. Additionally, employing BiRefNet for background removal enhances the quality of interactive content compared to alternatives like Segment Anything.

Research Question 3: How to minimize the potential latency during the interaction pipeline?

Minimizing latency is crucial to ensure real-time interaction and responsiveness in interactive art experiences. To achieve this, we explore various tools and techniques. We modify our interactive content from motion-to-motion/animation to motion-to-flower due to model performance constraints. This change allows us to leverage Stable Diffusion for its superior performance in generating images efficiently within the desired timeframe. The implementation of LCM-LoRA accelerates the inference process, reducing the time required for image generation. Additionally, BiRefNet stands out as an efficient solution for background removal, ensuring a streamlined input for the generative models. Through rigorous performance testing, we evaluate the efficiency of different tools and configurations, prioritizing those that provide a balance between speed and image quality. This approach ensures that the interactive system can respond promptly to user inputs, maintaining an immersive and engaging experience.

6.2 Limitation

Our project encountered several limitations that impacted the overall effectiveness and user experience, such as participants' background, testing environment, user-reported limitations, design limitations and GenAI flexibility.

One example of limitations regarding the user test is the background of participants. Most participants were young people with backgrounds in computer science, machine learning, VR, AR, and engineering. They are experienced with software design, human computer interaction and AI. However, it's worth noting that this demographic may not represent a fair cross-section of potential users. This technical expertise and specific age group may have influenced their interaction with the system and their feedback, potentially introducing bias.

Another limitation is the testing environment. The initial testing environment was an immersive cube, equipped with multiple cameras providing high-quality motion capture and higher frame rates. However, due to the renovation of the venue, subsequent testing was conducted using a big screen and lower quality motion capture technology. This resulted in reduced motion capture accuracy and sense of immersion. The new testing environment limited the user experience and the system's performance.

Several design limitations also affected user interaction and experience. The absence of on-screen visual effect of hand movements meant users lacked visual cues to indicate hand positions, which could have assisted in more precise interactions. Additionally, the system lacked adequate multi-sensory feedback, such as audio, which could enhance the immersive experience and help users understand their interactions better.

The flexibility of the GenAI prompts is currently limited. The prompts are predefined to only generate flowers, although the types and colors of the flowers are random. This limits the variety of creative outputs that the AI can produce, potentially reducing user engagement and exploration.

6.3 Future Work

Addressing the limitations identified before is essential for improving the system's performance and user experience. Future work will focus on following several key areas.

To obtain a more comprehensive understanding of user needs and system performance, future testing will involve a more diverse participant pool. This includes individuals of different ages, backgrounds, and experience. More participants would help ensure that the system is accessible and enhance the credibility of the SUS.

Improving the testing environment is crucial for achieving more reliable motion capture and a more immersive experience. We plan to test in immersive cube environment, using multiple cameras and higher frame rates for stable and accurate motion capture. This also avoids multiple calibrations. Additionally, developing a program to accurately measure the latency of hand movement visualization in Unity during testing will facilitate identification of latency sources and enhance real-time performance.

Several design improvements will be implemented to make the system more user-friendly and intuitive. Clear instructions and tutorials will be developed to guide users on understanding and interacting with the system effectively. On-screen hand tips will be introduced to provide visual cues for hand positions. An undo function can improve system's ability of error tolerance. Additionally, incorporating multi-sensory feedback, such as audio, will enhance the immersive experience and help users understand their interactions better.

The flexibility of the GenAI prompts will be explored to allow for more diverse and interactive content creation. This will include integrating other technologies, such as voice input, to enable users to generate a wide variety of elements.

Efforts will be made to reduce system latency and improve real-time responsiveness. This involves optimizing the motion capture system and refining algorithms to ensure smooth and accurate interactions.

By addressing these areas, we aim to create a more user-friendly, and engaging interactive system. These improvements will enhance the overall user experience, making the system more accessible and enjoyable for a diverse range of users, and paving the way for innovative applications in digital art and creativity.

7 Conclusions

In this chapter, we show some key findings from our development and user test. We also introduce the contributions on art creation, education, human computer interaction and society. Some recommendations are offered when using this project in practice.

7.1 Key Findings

Our project revealed several key findings regarding the integration of GenAI and motion capture technologies to create artistic interactive experiences.

Participants were satisfied with the generated flowers. They expressed a strong interest in using the system again. Participants emphasized the enjoyment they got from observing the flowers generated following the motions and creating art with friends. The collaboration facilitated communication, idea sharing, and creativity. The ability to work together makes the creative process more dynamic and enjoyable, highlighting the system's potential to enhance social interaction through art.

Real-time performance was identified as a crucial factor for user satisfaction. Participants were satisfied with the time taken by the GenAI to produce flowers, which we integrate with acceleration module. The latency between motions and the drawings was intolerable.

During exploring GenAI, we found some generated content occasionally lacked controllable quality, needing predefined prompts and picture depth information to ensure consistency. Additionally, the AI model exhibited biases due to the data on which it was trained, impacting the fairness and diversity of generated content.

7.2 Contributions

Our project shows how GenAI and motion capture technologies can change art creation, human computer interaction and education. By allowing users to create diverse and AI-generated content, our system offers a new way to teach and learn art. This hands-on approach makes learning more engaging and helps users understand artistic concepts. The unpredictable nature of the AI-generated content also encourages users to experiment and be creative.

The collaborative features of our system were a key highlight for users. They enjoyed drawing with friends and seeing flowers generated along their drawn lines. The system encourages communication and cooperation, allowing users to share the creative process. Working together helps people use each other's strengths and produce more varied art. This shared experience not only boosts individual creativity but also builds a sense of community and joint achievement.

The combination of GenAI and motion capture in interactive art has broader effects on society. It makes art creation accessible to more people, regardless of their artistic skills or experience. This inclusivity can inspire people from different backgrounds to engage in art, promoting cultural and creative expression. The collaborative and interactive nature of the system can also help strengthen social bonds and encourage collective creativity, which is important in today's digital age.

In conclusion, our project not only advances interactive digital art but also has significant potential for improving art education, fostering collaboration and creativity, and benefiting society. By continuing to

refine and expand our system, we aim to create an even more inclusive and impact tool for artistic and educational purposes.

7.3 Recommendations

To enhance the performance and user experience of our interactive system, several recommendations are proposed. High-performance GPUs are crucial to ensure smooth and efficient processing of both motion capture data and GenAI. This will help reduce latency and improve the responsiveness of the system. Employing high-quality motion capture systems providing accurate and real-time data is also essential. Systems with multiple cameras and high frame rates can minimize latency and improve accuracy of capturing users' motions.

Implementing the system in immersive environments, such as LBE setups or immersive cubes, is recommended to enhance user engagement. Regular checks and adjustments to keep the system well-calibrated is necessary.

When changing AI models and generating content, exploring suitable prompts and models is necessary. It is important to determine the specific information required by the GenAI to produce high-quality and stable content. Identifying biases in the GenAI models is also important to ensure fair and inclusive content generation.

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. “Gpt-4 technical report”. *arXiv preprint arXiv:2303.08774* (2023).
- [2] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras. “An integrated platform for live 3d human reconstruction and motion capturing”. *IEEE Transactions on Circuits and Systems for Video Technology* **27**:4 (2016), pp. 798–813.
- [3] C. Anthes, R. J. García-Hernández, M. Wiedemann, and D. Kranzlmüller. “State of the art of virtual reality technology”. In: *2016 IEEE aerospace conference*. IEEE. 2016, pp. 1–19.
- [4] M. A. Bagiwa, A. W. A. Wahab, M. Y. I. Idris, S. Khan, and K.-K. R. Choo. “Chroma key background detection for digital video using statistical correlation of blurring artifact”. *Digital Investigation* **19** (2016), pp. 29–43.
- [5] R. Barmaki and C. E. Hughes. “Providing real-time feedback for student teachers in a virtual rehearsal environment”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015, pp. 531–537.
- [6] P. R. A. S. Bassi, S. S. J. Dertkilgil, and A. Cavalli. “Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization”. *Nature Communications* **15**:1 (Jan. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-023-44371-z. URL: <http://dx.doi.org/10.1038/s41467-023-44371-z>.
- [7] E. Brynjolfsson, D. Li, and L. R. Raymond. *Generative AI at work*. Tech. rep. National Bureau of Economic Research, 2023.
- [8] J. Carmigniani and B. Furht. “Augmented reality: an overview”. *Handbook of augmented reality* (2011), pp. 3–46.
- [9] E. Catmull and R. Rom. “A class of local interpolating splines”. In: *Computer aided geometric design*. Elsevier, 1974, pp. 317–326.
- [10] F.-A. Croitoru, V. Hondu, R. T. Ionescu, and M. Shah. “Diffusion models in vision: a survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [11] Y.-C. Du, S.-C. Fan, and L.-C. Yang. “The impact of multi-person virtual reality competitive learning on anatomy education: a randomized controlled study”. *BMC Medical Education* **20** (2020), pp. 1–10.
- [12] Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, et al. “Art and the science of generative ai”. *Science* **380**:6650 (2023), pp. 1110–1111.
- [13] A. Escamilla, J. Meléndez, C. Monzo, and J. A. Morán. “Interaction designers’ perceptions of using motion-based full-body features”. *International Journal of Human-Computer Studies* **155** (2021), p. 102697.
- [14] M. Feng, A. Dey, and R. W. Lindeman. “An initial exploration of a multi-sensory design space: tactile support for walking in immersive virtual environments”. In: *2016 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. 2016, pp. 95–104.
- [15] G. Ferrand, J. English, and P. Irani. “3d visualization of astronomy data cubes using immersive displays”. *arXiv preprint arXiv:1607.08874* (2016).
- [16] G. Giarmatzis, E. I. Zacharaki, and K. Moustakas. “Real-time prediction of joint forces by motion capture and machine learning”. *Sensors* **20**:23 (2020), p. 6933.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. *Advances in neural information processing systems* **27** (2014).

- [18] A. Guo, X. Chen, H. Qi, S. White, S. Ghosh, C. Asakawa, and J. P. Bigham. “Vizlens: a robust and interactive screen reader for interfaces in the real world”. In: *Proceedings of the 29th annual symposium on user interface software and technology*. 2016, pp. 651–664.
- [19] C. Guo, X. Zuo, S. Wang, and L. Cheng. “Tm2t: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 580–597.
- [20] F. Hafiz, A. Shafie, O. Khalifa, and M. Ali. “Foreground segmentation-based human detection with shadow removal”. In: *International Conference on Computer and Communication Engineering (ICCCE’10)*. IEEE. 2010, pp. 1–6.
- [21] T. Hák, S. Janoušková, and B. Moldan. “Sustainable development goals: a need for relevant indicators”. *Ecological indicators* **60** (2016), pp. 565–573.
- [22] N. U. Islam and J. Park. “Depth estimation from a single rgb image using fine-tuned generative adversarial network”. *IEEE Access* **9** (2021), pp. 32781–32794.
- [23] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. “Motiongpt: human motion as a foreign language”. *Advances in Neural Information Processing Systems* **36** (2024).
- [24] M. Kaufmann, V. Vechev, and D. Mylonopoulos. July 2022. DOI: 10.5281/zenodo.1234. URL: <https://github.com/eth-ait/aitviewer>.
- [25] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. “Transformers in vision: a survey”. *ACM computing surveys (CSUR)* **54**:10s (2022), pp. 1–41.
- [26] D. P. Kingma and M. Welling. “Auto-encoding variational bayes”. *arXiv preprint arXiv:1312.6114* (2013).
- [27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. “Segment anything”. *arXiv:2304.02643* (2023).
- [28] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuohori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer. “Streamdiffusion: a pipeline-level solution for real-time interactive generation” (2023). arXiv: 2312.12491 [cs.CV].
- [29] F. Lamberti, G. Paravati, V. Gatteschi, A. Cannavo, and P. Montuschi. “Virtual character animation based on affordable motion capture and reconfigurable tangible interfaces”. *IEEE transactions on visualization and computer graphics* **24**:5 (2017), pp. 1742–1755.
- [30] P. Langdon and H. Thimbleby. *Inclusion and interaction: Designing interaction for inclusive populations*. 2010.
- [31] J. R. Lewis. “The system usability scale: past, present, and future”. *International Journal of Human–Computer Interaction* **34**:7 (2018), pp. 577–590.
- [32] Y.-L. Li, X. Wu, X. Liu, Y. Dou, Y. Ji, J. Zhang, Y. Li, J. Tan, X. Lu, and C. Lu. “From isolated islands to pangea: unifying semantic space for human action understanding”. *arXiv preprint arXiv:2304.00553* (2023).
- [33] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. “Gligen: open-set grounded text-to-image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22511–22521.
- [34] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*. Jan. 2024. URL: <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. “Smpl: a skinned multi-person linear model”. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866.
- [36] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao. “Lcm-lora: a universal stable-diffusion acceleration module”. *arXiv preprint arXiv:2311.05556* (2023).
- [37] P. Lyon, P. Letschka, T. Ainsworth, and I. Haq. “An exploratory study of the potential learning benefits for medical students in collaborative drawing: creativity, reflection and ‘critical looking’”. *BMC medical education* **13** (2013), pp. 1–10.
- [38] S. Macdonald. “Interconnecting: museum visiting and exhibition design”. *CoDesign* **3**:S1 (2007), pp. 149–162.
- [39] R. L. Mandryk, M. S. Atkins, and K. M. Inkpen. “A continuous and objective evaluation of emotional experience with interactive play environments”. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2006, pp. 1027–1036.
- [40] V. A. Mateescu and I. V. Bajić. “Attention retargeting by color manipulation in images”. In: *Proceedings of the 1st International Workshop on Perception Inspired Video Processing*. 2014, pp. 15–20.

- [41] J. Nielsen. “Ten usability heuristics” (2005).
- [42] J. F. O’Brien, R. E. Bodenheimer, G. J. Brostow, and J. K. Hodgins. “Automatic joint parameter estimation from magnetic motion capture data”. *arXiv preprint arXiv:2303.10532* (2023).
- [43] B. P. Ortega and J. M. J. Olmedo. “Application of motion capture technology for sport performance analysis”. *Retos: nuevas tendencias en educación física, deporte y recreación* 32 (2017), pp. 241–247.
- [44] S. Pargaonkar. “A comprehensive review of performance testing methodologies and best practices: software quality engineering”. *International Journal of Science and Research (IJSR)* **12**:8 (2023), pp. 2008–2014.
- [45] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf. *Diffusers: State-of-the-art diffusion models*. <https://github.com/huggingface/diffusers>. 2022.
- [46] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. “Sdxl: improving latent diffusion models for high-resolution image synthesis”. *arXiv preprint arXiv:2307.01952* (2023).
- [47] S. S. Raseli, N. A. M. K. Faisal, and N. Mahat. “The construction of cubic bezier curve”. *Journal of Computing Research and Innovation* 7:2 (2022), pp. 111–120.
- [48] P. A. Rauschnabel, R. Felix, C. Hinsch, H. Shahab, and F. Alt. “What is xr? towards a framework for augmented and virtual reality”. *Computers in human behavior* **133** (2022), p. 107289.
- [49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.
- [50] L. Senol, H. Gecili, and P. O. Durdu. “Usability evaluation of a moodle based learning management system”. In: *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE). 2014, pp. 850–858.
- [51] C. Solomon and T. Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.
- [52] R. Srinivasan and K. Uchino. “Biases in generative art: a causal look from the lens of art history”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 41–51.
- [53] B. C. Stahl and D. Wright. “Ethics and privacy in ai and big data: implementing responsible research and innovation”. *IEEE Security & Privacy* **16**:3 (2018), pp. 26–33.
- [54] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. “Motionclip: exposing human motion generation to clip space”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 358–374.
- [55] G. Tsampounaris, K. El Raheb, V. Katifori, and Y. Ioannidis. “Exploring visualizations in real-time motion capture for dance education”. In: *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. 2016, pp. 1–6.
- [56] L. van Velsen, T. van der Geest, and R. Klaassen. “Testing the usability of a personalized system: comparing the use of interviews, questionnaires and thinking-aloud”. In: *2007 IEEE International Professional Communication Conference*. IEEE. 2007, pp. 1–8.
- [57] R. V. Vitali and N. C. Perkins. “Determining anatomical frames via inertial motion capture: a survey of methods”. *Journal of Biomechanics* **106** (2020), p. 109832.
- [58] A. D. Wilson. “Touchlight: an imaging touch screen and display for gesture-based interaction”. In: *Proceedings of the 6th international conference on Multimodal interfaces*. 2004, pp. 69–76.
- [59] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. “Diffusion models: a comprehensive survey of methods and applications”. *ACM Computing Surveys* **56**:4 (2023), pp. 1–39.
- [60] M. N. H. Yunus, M. H. Jaafar, A. S. A. Mohamed, N. Z. Azraai, and M. S. Hossain. “Implementation of kinetic and kinematic variables in ergonomic risk assessment using motion capture simulation: a review”. *International Journal of Environmental Research and Public Health* **18**:16 (2021), p. 8342.
- [61] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen. “T2m-gpt: generating human motion from textual descriptions with discrete representations”. *arXiv preprint arXiv:2301.06052* (2023).
- [62] L. Zhang, A. Rao, and M. Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023.

- [63] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe. “Bilateral reference for high-resolution dichotomous image segmentation”. *arXiv* (2024).
- [64] E. Zhou and D. Lee. “Generative artificial intelligence, human creativity, and art”. *PNAS Nexus* **3**:3 (Mar. 2024), pgae052. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae052. eprint: <https://academic.oup.com/pnasnexus/article-pdf/3/3/pgae052/57464715/pgae052.pdf>. URL: <https://doi.org/10.1093/pnasnexus/pgae052>.
- [65] W. Zhou, Y. E. Jiang, E. Wilcox, R. Cotterell, and M. Sachan. “Controlled text generation with natural language instructions”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 42602–42613.

A Appendix

A.1 Questionnaires for the user test

Table A.1 Questions for the User Information Questionnaire

| |
|---|
| Q1. Email: [text] |
| Q2. Name: [text] |
| Q3. Gender: [Multiple choice] |
| Q4. Age: [text] |
| Q5. Work/Study field: [text] |
| Q6. Experience with Interaction Technologies: [Checkbox (Multi-choice)] Big Screen Interaction (e.g., interactive displays, touchscreens) Gesture-based Interfaces (e.g., Kinect, Leap Motion) Virtual Reality, Augmented Reality Motion Capture None Other: [text] |
| Q7. How often do you interact with interaction technologies in your daily life? [Checkbox (Single-choice)] Never Rarely Occasionally Frequently Almost Constantly |
| Q8. Experience with AI Interaction Modalities: [Checkbox] Text-based AI Interaction (e.g., chatbots, virtual assistants) Image Generation AI (e.g., Gencraft, Shutterstock) Sound Generation AI (e.g., Speechify - Voice generator) AI-driven Personalization (e.g., content recommendations, personalized user experiences) None Other: [text] |
| Q9. How often do you use AI Interaction Modalities? [Checkbox (Single-choice)] Never Rarely Occasionally Frequently Almost Constantly |
| Q10. What are your expectations for a project that combines human motion with generative AI to create artistic works? [text] |

Table A.2 Questions for the Project Evaluation Questionnaire

| Exploration Task |
|--|
| Q1. Did you find the interaction with the system intuitive and easy to understand? [5-point Likert scale] |
| Q2. How expressive did you feel your hand motions were in influencing the size and shape of the generated flowers? [5-point Likert scale] |
| Q3. How much did you enjoy experimenting with using hand motions to create flower images? [5-point Likert scale] |
| Q4. Overall, how satisfied are you with the exploration task? [5-point Likert scale] |
| Q5. Were there any challenges or difficulties you encountered while exploring different hand motions? [text] |
| Standardized Task |
| Q6. How would you rate the ease and accuracy of drawing lines using the motion capture system? [5-point Likert scale] |
| Q7. How satisfied were you with the flower images generated from your line drawings? [5-point Likert scale] |
| Q8. How satisfied were you with the response time from drawing a line to seeing the generated image? [5-point Likert scale] |
| Q9. Did you find the addition of the line smoothing function beneficial when generating flower images from your drawings? [5-point Likert scale] |
| Q10. How visually appealing did you find the flower images generated from your drawings? [5-point Likert scale] |
| Q11. How closely did the generated flower images resemble the shapes you drew? (size, position, shape) [5-point Likert scale] |
| Q12. How much did you enjoy the standardized task of creating flower images? [5-point Likert scale] |
| Q13. How interested would you be in trying this interactive experience in the future? [5-point Likert scale] |
| Q14. Please describe any challenges or difficulties you encountered while drawing different lines. [text] |
| Q15. Do you have any suggestions or feedback for improving the system, such as additional features or enhancements? [text] |
| Co-creation Task |
| Q16. How much did you enjoy the collaborative creation of flower images? [5-point Likert scale] |
| Q17. How would you rate the ease of the collaborative creation of flower images? [5-point Likert scale] |
| Q18. Which creation experience did you enjoy more? [5-point Likert scale] |
| Q19. What aspects of the interaction experience did you find most enjoyable or interesting? [text] |
| Q20. How would you describe your experience while exploring different hand motions to generate flower images? [text] |

Table A.3 Questions for the System Usability Scale Questionnaire

| |
|--|
| Q1. I think that I would like to use this system frequently. [5-point Likert scale] |
| Q2. I found the system unnecessarily complex. [5-point Likert scale] |
| Q3. I thought the system was easy to use. (easily enough with the interface to complete tasks/goals effortlessly) [5-point Likert scale] |
| Q4. I think that I would need the support of a technical person to be able to use this system. [5-point Likert scale] |
| Q5. I found the various functions in this system were well integrated. [5-point Likert scale] |
| Q6. I thought there was too much inconsistency in this system. [5-point Likert scale] |
| Q7. I would imagine that most people would learn to use this system very quickly. [5-point Likert scale] |
| Q8. I found the system very cumbersome to use. [5-point Likert scale] |
| Q9. I felt very confident using the system. [5-point Likert scale] |
| Q10. I needed to learn a lot of things before I could get going with this system. [5-point Likert scale] |