LATENT AUTO-RECURSIVE COMPOSITION ENGINE: A GENERATIVE SYSTEM FOR CREATIVE EXPRESSION IN HUMAN-AI COLLABORATION

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Master of Science

in

Computer Science

by Yenkai Huang

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

May 2024

Examining Committee:

Michael A. Casey, Chair

Lorie Loeb

Elizabeth L. Murnane

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

This thesis investigates the shifting boundaries of art in the era of Generative AI, critically examining the essence of art and the legitimacy of AI-generated works. Despite significant advancements in the quality and accessibility of art through generative AI, such creations frequently encounter skepticism regarding their status as authentic art. To address this skepticism, the study explores the role of creative agency in various generative AI workflows and introduces an "artist-in-the-loop" system tailored for image generation models like Stable Diffusion. This system aims to deepen the artist's engagement and understanding of the creative process. Additionally, a novel tool, the Latent Auto-recursive Composition Engine (LACE), which integrates Photoshop and ControlNet with Stable Diffusion, is introduced to improve transparency and control. This approach not only broadens the scope of computational creativity but also enhances artists' ownership of AI-generated art, bridging the divide between AI-driven and traditional human artistry in the digital landscape.

Acknowledgments

I extend my deepest gratitude to my thesis advisors—Michael A. Casey, James Mahoney, Michael Cohen, and Elizabeth L. Murnane—for their invaluable guidance and support and for providing me with an inspiring research environment. Their expertise and mentorship have been pivotal in shaping both the direction and success of my study.

I am particularly thankful to Lorie Loeb for welcoming me into this program and introducing me to the fascinating world of computer science. Her belief in my potential has been a constant source of motivation.

My journey through the MSDA and MSCS cohorts has been enriched and supported by a wonderful community of peers. I am especially grateful to my close companions—Hadid, Soo, Diana, and Jasper—with whom I have shared countless late nights, tirelessly working towards our common academic goals. Their camaraderie and emotional support have been a cornerstone of my graduate experience.

I must also acknowledge Ravi's technical prowess, which came to the rescue of my training models by fixing the continuous issues with my PC setup. His help was crucial in keeping my research on track.

I would like to express my sincere appreciation to all the participants in my study. Their involvement and the experiences they shared were essential to the depth and breadth of my research. Their contributions have been invaluable, and I am deeply grateful for their participation.

A special thanks to YiKang, whose guidance was instrumental in my application process. Her advice and support helped pave the way for my acceptance into this esteemed program, and I am profoundly grateful for her assistance.

Lastly, I owe a profound thank you to my wife, Grace, and my family. Their endless patience and emotional support have sustained me through the intense and often challenging moments of my academic journey. Their ability to endure my endless tech-related monologues, reminiscent of Gollum's obsessions, has been nothing short of heroic.

This thesis would not have been possible without the support and encouragement of each individual mentioned, and I am immensely thankful for their contributions to my academic and personal growth.

Preface

Every generation is defined by its relationship with the media and technologies of its time. Currently, artificial intelligence (AI) represents such a frontier, not only as a field of scientific endeavor but also as a medium of creative expression. As an artist and designer, I am fascinated yet often perplexed by the design of AI systems, which seldom reflect the artist's mindset. Despite this, the allure of AI is undeniable, evoking a modern-day "Lord of the Rings," tempting every tech enthusiast to explore its capabilities.

This thesis emerges from a pivotal question in contemporary art discourse: Can AI-generated works be legitimately considered art? Rather than merely replicating existing artistic styles, this work investigates how AI can expand our creative horizons. While prevailing approaches in computational creativity are predominantly algorithmic, they frequently lack integration with the artistic perspective. My research aims to reconcile these perspectives by fostering a symbiosis between technological innovation and artistic intuition.

To tackle these issues, it is essential to critically analyze the concept of art and the artistic process, particularly through the lens of human-centered computing. Adopting a Human-Computer Interaction (HCI) perspective, this thesis explores various workflows and proposes innovative methodologies designed to bridge the gap between generative AI technology and traditional artistic practices.

The core of this research focuses on addressing the unpredictability and limited control inherent in AI-generated outputs, challenges that are particularly pronounced in diffusion-based AI models. These concerns are not merely theoretical but impact the practical utility of AI in artistic workflows.

This document not only serves as a record of my scholarly and creative journey but also as a demonstration of the potential for AI tools to be integrated meaningfully into artistic practices. The workflows developed and discussed herein offer a new framework for artists and researchers eager to navigate the evolving landscape at the intersection of art and technology. Through this exploration, I aim to demonstrate that AI, when thoughtfully integrated, can enrich the creative process, extending the boundaries of what is possible in the realm of art. By blending advanced technology with traditional artistic methods, this thesis illustrates how AI can serve not merely as a tool, but as a dynamic collaborator in the creative process. My goal is to foster a deeper understanding and appreciation of AI's role in art, inspiring both artists and technologists to pursue innovations that enhance and expand the artistic landscape for future generations.

Contents

Chapter 1

Background and Prior Work

Section 1.1

The Essence of Art

The capability of generative AI to create art has been a topic of scholarly debate for decades. Visually, the output of these AI models is very successful. However, some might argue that AI is struggling with generating contextual meaning and the innovative structure of the art piece. Moreover, the art community doesn't regard the output as a work of art, thinking the statistical learning from previous work is just a copycat of humans's endeavor. The inherent subjectivity in evaluating art, particularly in the field of computational creativity, poses a significant challenge in assessing machine-generated works. This issue is further complicated by the lack of a universal standard for measuring artistic quality.

To meaningfully advance this discussion, it is crucial to establish a clear definition of what constitutes art. This foundational step will provide a framework for developing a well-grounded thesis that addresses the complexities of computational creativity. By defining the boundaries and characteristics of art, we can more effectively assess the artistic merit of computer-generated works and explore the various dimensions of this emerging field.

1.1.1. What Makes an Art

Discussions surrounding the definition of art have generated plentiful theoretical frameworks, yet no theory has successfully encompassed all aspects of art. Philosophers such as Morris Weitz have even challenged the pursuit of defining art's essence, arguing that art is an inherently "open concept"[43]

However, adopting a broader definition may be beneficial. George Dickie's perspective in "Defining art"[18] offers a foundational approach to identifying what qualifies as a work of art. Dickie differentiates between the generic concept of "art" and specific sub-concepts like novels, tragedies, or paintings. He suggests that while these sub-concepts may not have the necessary and sufficient conditions for definition, the overarching category of "art" can be defined.

According to Dickie, two key elements are essential: a) **artifactuality**—being an artifact created by humans; and b) the **conferring of status**—where a society or subgroup thereof has recognized the item as a candidate for appreciation.

This framework seeks to avoid the pitfalls of traditional art definitions, which often implicitly include notions of "good art," are overly restrictive, or depend on metaphysical assumptions. Instead, his definition aims to reflect the actual social practices within the art world. [15]

Building on this legacy, this thesis will explore artifactuality through computational methods in the context of generative AI, focusing on one of the foundational elements that might define what can be considered art in the digital age.

1.1.2. Can Humans Distinguish Between Human and Machine-made Art?

Addressing George Dickie's concept of the "conferring of status," a pivotal inquiry arises: can people discern between artworks created by humans and those generated by machines, especially when the quality of machine-made art rivals that of human creation?

The 2024 study by Kazimierz Rajnerowicz [34] reveals a growing difficulty in distinguishing between AI-generated and human-created images, with up to 87% of participants unable to make accurate identifications, and this difficulty persists even among those with AI knowledge. Rajnerowicz's article examines how individuals judge the authenticity of images, the potential risks of failing to recognize AI-generated content, and underscores the necessity of understanding AI advancements to prevent deception by deepfakes and other sophisticated AI techniques.

Lucas Bellaiche et al.[7], delves into the perception and contextual meaning between humans and AI-generated art. Their study indicates that people tend to perceive art as reflecting a human-specific experience, though creator labels seem to mediate the ability to derive deeper evaluations from art. Thus, creative products like art may be achieved—according to human raters—by non-human AI models, but only to a limited extent that still protects a valued anthropocentrism.

However, a prospect explored by Demmer and colleagues in their study, "Does an emotional connection to art really require a human artist?" [16] Their study uncovered compelling evidence indicating that participants experienced emotions and attributed intentions to artworks, independent of whether they believed the pieces were created by humans or computers. This finding challenges the assumption that AI-generated art is incapable of evoking emotional and intentional human elements, as participants consistently reported emotional responses even towards computer-generated images.

Nonetheless, the origin of the artwork did have an impact, with creations by human artists eliciting stronger reactions and viewers often recognizing the intended emotions by the human artists, suggesting a nuanced perception influenced by the actual provenance of the art.

1.1.3. Why do people think AI generative artwork is “artificial”?

Generative AI empowers artists to manipulate the latent space with ease, creating new artworks through simple prompt modifications. Advanced techniques and tools such as LoRa[24], ControlNet[47], and image inpainting[5] provide even greater control over the generative output. Despite these capabilities, there is a prevalent bias among observers who view computer-generated art as “artificial” [9]. This perception stems from several factors:

- **Lack of Human Touch**: AI-generated art lacks a direct human creative process, leading to views of it being less genuine or lacking soul.
- **Reproducibility**: The ability of AI to rapidly produce multiple, similar outputs may reduce the perceived value and uniqueness of each artwork.
- **Transparency and Understanding**: The opaque decision-making process of AI systems often results in doubts about the creativity involved [9].
- **Missing Context**: AI does not fully understand or express the social, cultural, or political nuances that deepen traditional art, often making its products seem technically proficient yet shallow.

1.1.4. Creative Process and Iterative Intent

In his 1964 work, “The Artworld,”[15] Arthur Danto emphasizes that art depends on an “artworld” consisting of theories, history, and conventions that recognize it as art rather than mere objects. This notion highlights a pivotal idea: in modern art, the creative process might be more important than the actual artwork. Without the artworld’s narratives and contexts, the audience may struggle to grasp the artwork since it is not guaranteed to deliver the same experience, potentially widening the gap between art perception and concept. Furthermore, this gap expands further in

AI-generated art due to the opaque nature of AI models, which obscure the creative process and question the legitimacy of the artwork.

Artists find themselves unable to articulate the relationship between their input (text prompts) and the machine-generated output, reducing their sense of ownership over the work. Despite potentially high-quality results, artists might not view these outputs as their own creations, leading to a perception of AI models as the true authors.

Moreover, the iterative nature of creative intent presents further challenges. Artists typically do not start with a clear vision; instead, they develop and refine their goals through the creative process, an approach fundamental in fields such as design, architecture, or illustration, where concepts often evolve through iterative experimentation. For instance, in architectural design, a technique known as "Generative drawing" plays a critical role. Described in "Generative Processes: Thick Drawing" [39], by Karl Wallick, this method involves using drawings not just as tools for documentation but as active participants in the design process. These drawings help conceptualize ideas while integrating both abstract thought and practical execution into a single visual narrative, maintaining visibility of the design process to enhance creative exploration.

This contrasts sharply with the requirements of most generative models, which necessitate a well-defined intent from users, usually articulated through precise text prompts. This rigid structure creates a significant disconnect: if an artist's intent shifts during the creative process—a common occurrence—the output from the generative model may no longer align with their evolving vision, rendering the quality of the result irrelevant.

Section 1.2

Sense of Agency

Agency in generative artwork refers to the capacity of the creator to make independent decisions that significantly affect the outcome of the art. In traditional art, agency is clear-cut; artists consciously choose every detail of their work, from the medium to the message. However, in AI-generated art, the concept of agency is more nuanced. The human operator provides text prompts or images, and the AI then processes these inputs based on its training data. Many artists contend that presenting the raw output of AI as one's own work amounts to theft and a lack of originality since these outputs are built on the contributions of other artists' styles, often utilized without consent concerning copyright and creative thought. Therefore, understanding the agency's complexities is crucial in human-AI collaborations, affecting the quality, ownership, and impact on computational creativity.

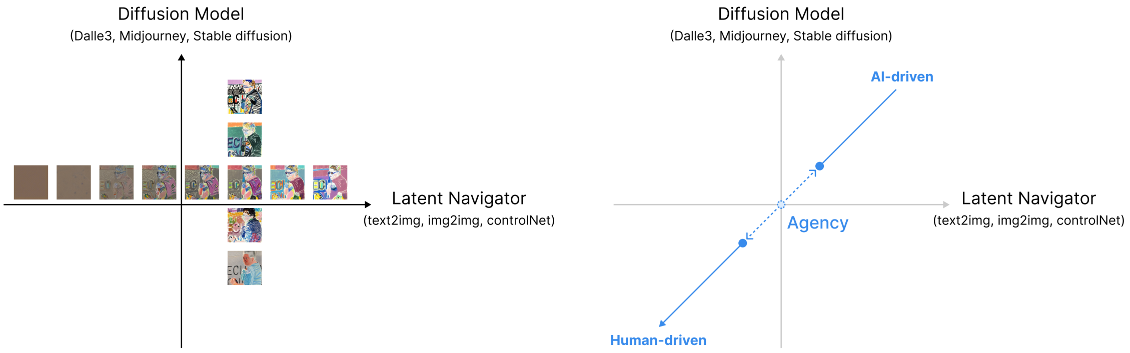

Diffusion Model
(Dalle3, Midjourney, Stable diffusion)
Latent Navigator
(text2img, img2img, controlNet)
Diffusion Model
(Dalle3, Midjourney, Stable diffusion)
AI-driven
Latent Navigator
(text2img, img2img, controlNet)
Agency
Human-driven

Figure 1.1: Charting artistic authenticity: A new metric for assessing the authenticity of artwork by mapping agency on an additional axis. Here, the spectrum of agency spans from fully autonomous AI-generated art to human-driven creation, offering a nuanced view of artistic origination.

1.2.1. The Importance of Agency

In HCI, studies on the Sense of Agency (SoA)[14], like Wegner et al.'s "Vicarious agency: experiencing control over the movements of others"[41] are prevalent to measure one's perception of the cause-and-effect on an event or object, and there for determine whether the person has ownership over it. However, these studies primarily focus on augmentation of body, action, or outcome rather than internal states like creativity aided by machines. Yet in art, particularly AI-enhanced creation, agency's significance extends to a proposed axis that gauges the artwork's authenticity.

AI art creation traditionally orbits two axes: the architecture grounding the model in training data, and the latent navigator that seeks the desired image from input prompts. This process epitomizes the creative journey in AI art, where the user steers the pre-trained model with tools to match their artistic vision.

Introducing the agency axis, depicted in Figure 1.1, reframes the artistic origin narrative. At one extreme, complete human intervention equates to a work that is entirely the artist's intellectual property, where every creative facet is handpicked. On the other hand, excessive reliance on AI for composition and style risks depersonalizing the art, prompting the art community's devaluation of such works.

An optimal creative balance is struck when both human and AI inputs interact, fostering a space where creative spontaneity meets deliberate artistry, and ownership over the end product is clear. This intersection becomes a breeding ground for creative discovery, marrying human intention with AI's potential.

1.2.2. Intentional Binding and Human-AI Interaction

A fundamental aspect of human cognition that illuminates our interaction with AI in creative processes is intentional binding [28]. This phenomenon, where individuals perceive a shorter time interval between a voluntary action and its sensory consequence, highlights how the perception of agency influences our engagement with the world. In the context of HCI, especially in AI-enhanced art, understanding intentional binding provides valuable insights into how artists perceive and integrate AI responses into their creative expression.

When artists interact with AI, the immediacy and relevance of the AI's output to the artist's input can affect their sense of control and creative ownership. If the output closely and quickly matches the artist's intention, similar to the effects observed in intentional binding, the artist may experience a greater sense of agency. This heightened perception of control can make AI tools feel more like an extension of the artist's own creative mind, rather than an external agent imposing its own logic.

Therefore, in discussions about agency in AI-generated art, it is crucial to consider how the principles of intentional binding might play a role in shaping the artist's experience of the creative process. This understanding can guide the development of more intuitive AI systems that enhance the artist's agency, promoting a more seamless and satisfying creative partnership.

Section 1.3 The Evolution of Image Synthesis

The art community has experienced profound transformations over the years, significantly influenced by advancements in artificial intelligence and machine learning. The evolution of generative models has been pivotal in shaping both the capabilities and applications of generative art. In this section, we will explore the history of technology in art.

1.3.1. Generative Models

The development of generative models in image synthesis and artistic creation has seen remarkable transformations since the mid-20th century. Initially, in the 1950s and 1960s, pioneering artists utilized oscilloscopes and analog machines to produce visual art directly derived from mathematical formulas. With the advent of more widely accessible computing technologies in the 1970s and 1980s, artists such as Harold Cohen began to employ these tools to create algorithmic art. Cohen's work with AAron [12], for instance, involved using programmed instructions to dictate the form and structure of artistic outputs. During the 1980s, the emergence of fractal art marked a significant advancement, employing mathematical visualizations to craft complex and detailed patterns. The 1990s introduced evolutionary art and interactive genetic algorithms, which further democratized the creative process by allowing both artists and viewers to participate in the evolution of artworks.

The resurgence of neural networks in the 2000s, fueled by advancements in GPU computing power and the availability of large datasets, significantly propelled technological innovation. Prominent developments such as Convolutional Neural Networks (CNNs)[19], Variational Autoencoders (VAEs)[26], and Generative Adversarial Networks (GANs)[22] transformed the landscape of generative art. These models revolutionized the field by generating highly realistic images that closely mimic the characteristics of their training data.

Recent advancements, such as Vision Transformers (ViTs) [31], Latent Diffusion Models (LDMs) [35], and Mixture of Experts (MoE) models like RAPHAEL [46], have pushed the boundaries by enabling the synthesis of complex images from detailed text prompts, facilitated by Contrastive Language-Image Pretraining (CLIP) [33]. These innovations merge artistic expression with cutting-edge technology and challenge traditional notions of creativity and the artist's role.

These developments highlight a progression from relatively simple predictive models to sophisticated systems capable of understanding and generating complex visual content, demonstrating the model can encode art concepts into embeddings and later reconstruct or synthesize them for new artistic purposes.

1.3.2. Control and interpretability

Control and interpretability in generative models have undergone transformative growth, enhancing both functionality and accessibility across various domains. This subsection explores key innovations in control mechanisms and interpretability within image synthesis and latent diffusion models.

Classifier-Free Diffusion Guidance is a technique used in generative diffusion models to enhance image quality without a separate classifier. Developed by Dhariwal and Nichol [23], it involves training the model to occasionally ignore the conditioning input, which allows flexibility during inference by adjusting the conditioning strength. This method efficiently guides the generation toward desired outputs more closely, enhancing the accuracy and diversity of the generated images without relying on additional discriminative components.

LoRA [24], or Low-Rank Adaptation, reduces the complexity of adapting large pre-trained models by introducing trainable low-rank matrices. This method significantly decreases the number of trainable parameters, enabling efficient fine-tuning of models such as Stable Diffusion for specialized styles or art concepts without extensive computational power or large datasets.

ControlNet [47] introduces spatial conditioning in pre-trained text-to-image models, allowing for precise manipulation of image elements based on multiple conditions such as edges, outlines, poses, and segmentation. This flexibility is instrumental in producing high-quality images tailored to specific requirements, making it a pivotal tool for artists.

Visual Prompting via Image Inpainting [6] presents a method for performing task-specific image transformations based on visual prompts. This approach utilizes masked auto-encoders trained on a unique dataset to generate task-consistent outputs, showcasing the model's adaptability to a range of visual tasks without extensive retraining.

The **Aesthetic Gradient** technique [20] personalizes image generation by aligning text prompts with user-defined aesthetic preferences. This method optimizes the text encoder of CLIP-conditioned models, steering the generative process towards desired visual styles with minimal computational overhead.

Tom White's exploration in **Sampling Generative Networks** [44] provides innovative techniques for sampling the latent spaces of models like VAEs and GANs. Methods such as spherical linear interpolation (slerp) enhance the quality of generated samples and facilitate the understanding of latent space dynamics.

Unsupervised Discovery of Semantic Latent Directions [30] introduces a novel method to explore and manipulate latent spaces in diffusion models. This approach utilizes Riemannian geometry to identify and utilize meaningful editing directions, enabling detailed control over attribute changes in generated images.

These advancements not only enhance the practicality and effectiveness of generative models but also contribute to a deeper understanding and accessibility of these technologies. They pave the way for more intuitive and user-centric applications, extending the reach of generative models beyond traditional boundaries and into more creative and personalized uses.

Section 1.4

The limitation of Text-to-Image Models

Text prompts serve as a versatile and universally accessible method to guide the generation of images. As large language models evolve, text-to-image models are increasingly capable of interpreting both literal and semantic meanings of text prompts. Nevertheless, these models often face challenges in accurately rendering complex or abstract concepts based solely on textual descriptions. This misalignment between the generated content and its intended semantics presents several issues that need addressing.

1.4.1. Misalignment of Text Encoding

Wu et al. (2023)[45] explore the disentanglement capabilities of stable diffusion models, demonstrating that these models can effectively differentiate between various image attributes. This disentanglement is facilitated by adjusting input text embeddings from neutral to style-specific descriptions during the later stages of the denoising process.

For instance, as illustrated in Figure 1.2 (Wu et al. 2022), the prompt "A photo of a woman" might yield significantly different results from "A photo of a woman with a smile." Although modifying text embeddings can help segregate different attributes, this approach struggles with fine, localized edits and may be overwhelmed by overly detailed neutral descriptions.

Moreover, dependency on prompt engineering often encounters inherent limitations as the model's semantic interpretation can significantly diverge from human understanding. Such dependence is typically restricted by the labels in the training dataset. For instance, using specialized terminology like "monogram" to describe a straightforward "lack and white" graphic may lead to unforeseen results, largely due

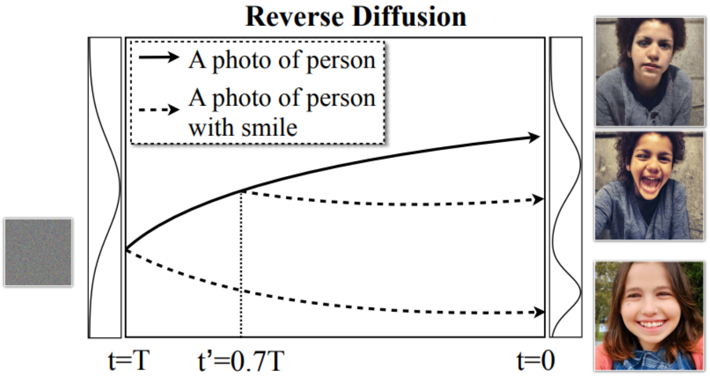


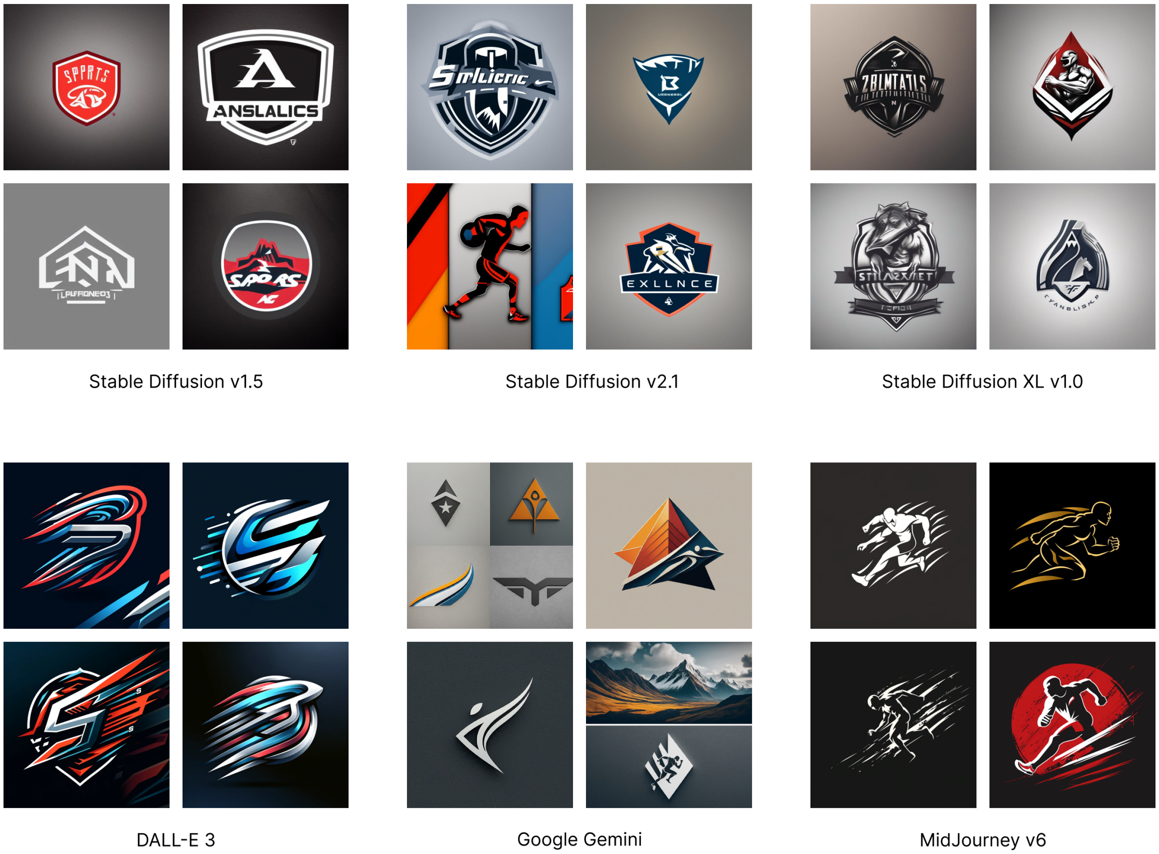
Figure 1.2: The disentanglement property of stable diffusion models. The top image is generated conditioned on “a photo of person”. The bottom image is generated with all descriptions replaced with “a photo of a person with smile”, and changes the person’s identity. The middle image is generated by partially replacing descriptions at later steps and maintaining the person’s identity. (Wu et al., 2022)

to the annotators’ limited domain-specific knowledge.

Despite potential improvements in semantic understanding and better alignment of text embeddings with image and art concepts as large image generation models become more complex and larger, users may still face difficulties in articulating abstract concepts through text. The main challenge arises from the discrepancy between how datasets are annotated and how users describe their desired outcomes using the same vocabulary.

1.4.2. Case Study: Logo Generation

An investigation into the limits of text encoding within generative models initiated a study focusing on the design potential of text-to-image models. This aimed to discern the AI’s capacity to generate novel design elements from abstract prompts. A key inquiry was whether AI could derive minimalist designs akin to Nike’s emblem from prompts like ”a logo for a sports brand emphasizing athletic excellence, innovation, and pushing performance boundaries,” or if it would default to creating literal representations similar to existing sports logos.



Stable Diffusion v1.5

Stable Diffusion v2.1

Stable Diffusion XL v1.0

DALL-E 3

Google Gemini

MidJourney v6

Figure 1.3: Comparison of minimalistic logo designs for a sports brand generated by state-of-the-art models. Prompt used: “Design a minimalistic logo for a sports brand emphasizing athletic excellence, innovation, and performance boundaries.”

Figure 1.3 illustrates a critical challenge in assessing AI-generated images for brand logo design. While the models succeeded in generating graphics with a sports theme, the nuanced symbolism and brand narratives, which are central to effective logo design, were not as clearly interpreted or presented. This underlines a significant gap in AI’s ability to embody abstract and culturally resonant design elements in its visual creations. In conventional design methodology, creatives often engage with clients to distill and agree upon symbolic elements that represent the brand. These elements serve as a foundation for the final logo that embodies the company’s intended message. This iterative and collaborative aspect of the design is absent in AI’s text-to-image process, leaving evaluators without a clear metric to determine the significance or quality of the generated logos.

To further investigate design tasks, I fine-tuned an existing model based on Stable Diffusion v1.5. by RunwayML [2]. The experiment aimed to determine the extent to which a text-to-image model could excel in generating logos without advancements in text encoding or refinement of semantic embedding alignments with the input prompts.

The dataset, "amazing_logos_v4" [25], was personally curated by me and consists of nearly 400,000 images of logos. These images were gathered from renowned design websites such as logo-archive.org and logolounge.com and are designated for non-commercial use. The researcher utilized these images to fine-tune the Stable Diffusion v1.5 model over 21 epochs, consuming approximately 500 GPU-hours on an Nvidia RTX 3090.

For data labeling, I designed a structured prompt that integrates the company's name, descriptive elements of the logo design, and the company's country and industry. This approach ensures that each text prompt effectively combines key design keywords with specific details pertinent to the company. The structure of the prompt is as follows:

```
Simple elegant logo for {company name},
{concept} {country}, {industry}, successful vibe, minimalist,
thought-provoking, abstract, recognizable
```

An example of such a prompt is:

```
Positive prompt:
    Simple elegant logo for Dartmouth College,
    D Pine tree circle United States, education, successful vibe,
    minimalist, thought-provoking, abstract, recognizable
Negative prompt:
    out of frame, low res, wooden background, collage
```

The study explored the model's proficiency in assimilating designated shapes into logo designs, a process that involved modifying input prompts to include shape descriptors like 'circle', 'square', 'wave', and 'triangle'. Observations from Figure 1.4 revealed a trend towards more minimalist logos as the model progressed through training epochs. Notably, the visual effectiveness varied with each shape and at different stages of training—'square' and 'dot' at epoch 15, and 'wave' and 'triangle' at epoch 18 showcased distinct design strengths.

Challenges mounted when the model was tasked with more abstract prompts. As depicted in Figure 1.5, judging the adequacy of responses to a prompt that required designing a digital art logo featuring 'D', 'A', 'computer', 'art', and 'circle' proved difficult. The indistinct context made it hard to confirm if the results resonated with the prompt's requirements. This uncertainty affirms that abstract symbolism demands a richer context for its interpretation.

In short, AI models variably integrated the specified shapes into logos, sometimes as central themes and occasionally as nuanced details. This illustrates the limitations of using text prompts for logo generation in AI, highlighting a shortfall in AI's capability to amalgamate new graphic elements with their associated social or cultural connotations from scant textual information. Despite advancements in language models such as ChatGPT-4 or Google Gemini and methodologies like chain of thought reasoning [42], AI systems continue to grapple with interpreting the literal versus the cultural context embedded within prompts.

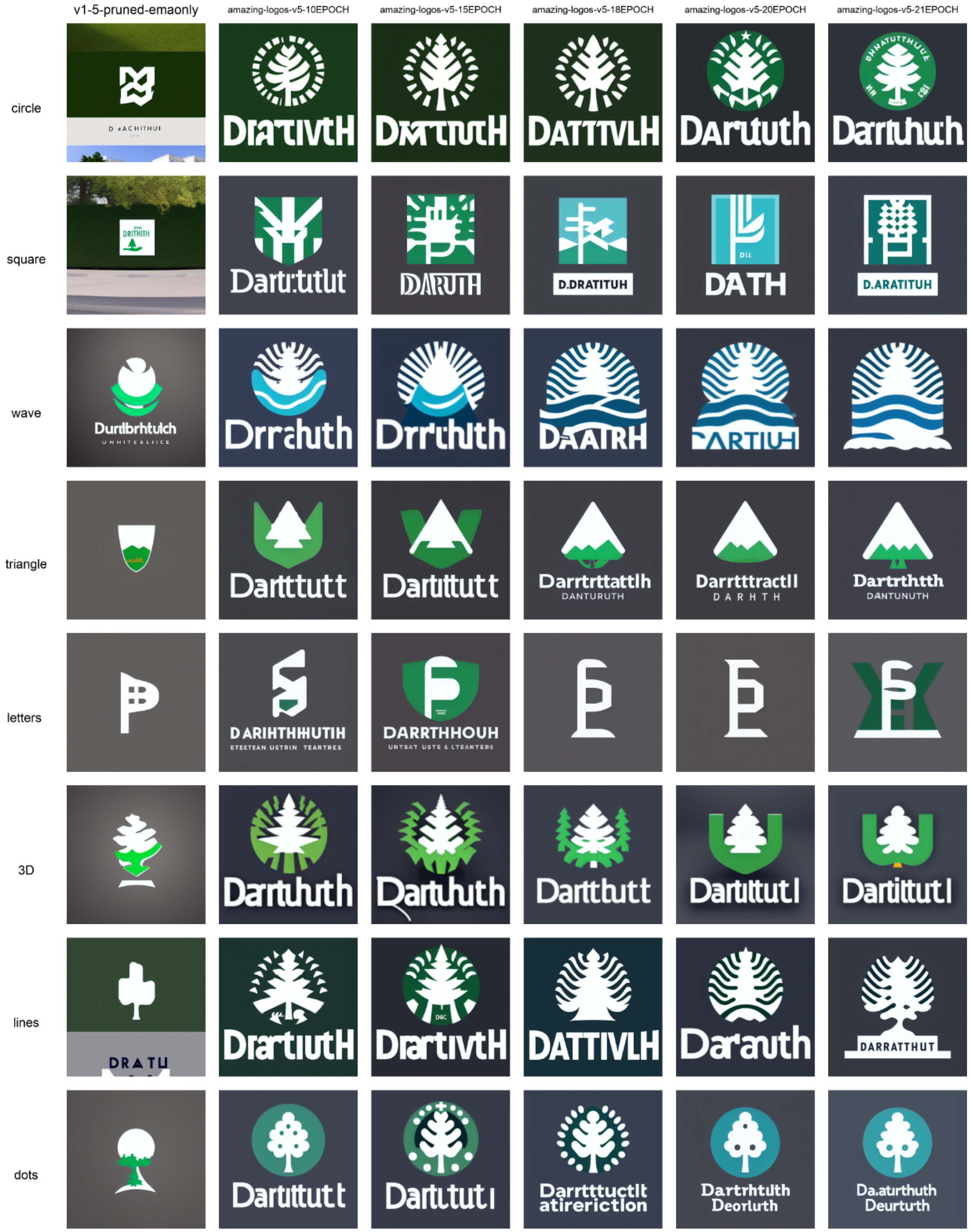


Figure 1.4: Comparison of logo generation from the prompt: "Simple elegant logo for Dartmouth College, D Pine tree circle United States, education, successful vibe, minimalist, thought-provoking, abstract, recognizable" The word 'circle' is replace by a list of words in y-axis

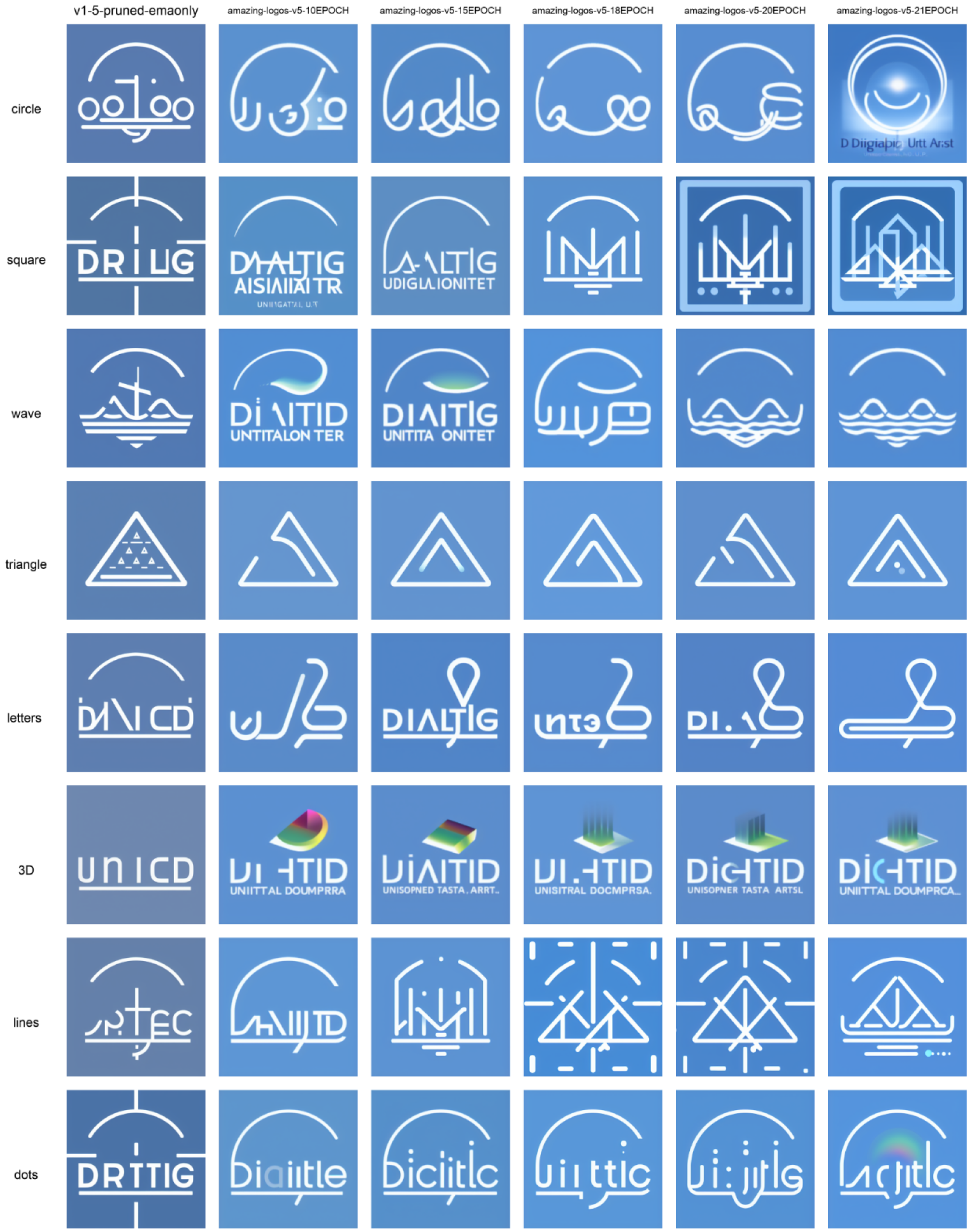


Figure 1.5: Comparison of logo generation from the prompt: “Simple elegant logo for digital art, D A computer art circle United states, education, successful vibe, minimalist, thought-provoking, abstract, recognizable” The word ‘circle’ is replace by a list of words in y-axis

Chapter 2

Integrating Human Creativity with AI in Art

Section 2.1

Analysis on Human-AI Collaborative Workflow

To gain a deeper understanding of the role of agency in generative art and ownership in the creative process, it is crucial to scrutinize the workflow of the creative process in text-to-image models. While this workflow could be adapted for other model types, the focus here is on image generation. This focus is chosen because visual outputs have a more direct correlation with the concept of art.

2.1.1. Workflows I: Iterative Prompt Engineering

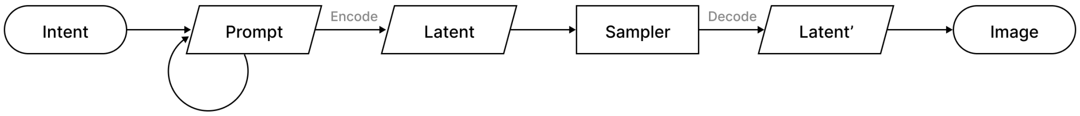


Figure 2.1: Users continuously update the text or image prompts

Iterative Prompt Engineering (Figure 2.1) is the prevalent method in text-to-image model workflows, where users continuously update the text or image prompts. These are encoded into latent representations and then sampled, with the results decoded back into images.

Considerable effort has been expended within the AI community to understand how generative models respond to various “magic words” to achieve desired outcomes. A prime example is the Stable Diffusion Prompt Book [17] by OpenArt, which provides an extensive guide on prompt formatting, parameters, artistic styles, and compositions to help users navigate the complexities of the model’s behavior.

Despite the potential effectiveness of prompt engineering, users often face challenges where slight variations—such as synonyms or changes in number (singular vs. plural)—can produce significantly different outcomes. Moreover, fine-tuning prompts typically require testing multiple keyword combinations, such as “art station trending,” “hi-res,” “4K,” or “masterpiece.” This process can be both time-consuming and counterintuitive. Additionally, whenever a new model is introduced—whether it features different training data or a new architecture—users must adapt anew to its distinct characteristics.

Regardless all the drawbacks, this workflow is still the most popular method adapted by most users. The rapid inference speeds of image generation models, as seen in LCM-based models [27] or platforms like Krea[1], provide a “live tweaking” experience, enhancing user agency. However, the inherent randomness of the sampling process and the complex interplay within the latent space can lead to unpredictable results.

2.1.2. Workflows II: Recursive Iteration in Latent Sampling Processes

The second workflow (Figure 2.2) type involves iterative modifications of the same latent vector to enhance control over the model. This approach involves sampling

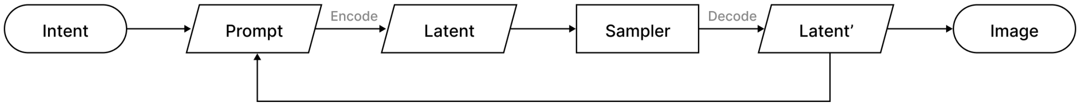


Figure 2.2: Users sampling the same latent vector repeatedly to refine the result

the same latent vector repeatedly; its embedding value evolves through each iteration due to the stochastic nature of the diffusion process.

Enhancing control over generative outcomes is achievable by refining the conditioning and workflow of the latent sampling process. For instance, using a previously sampled latent vector as an input in subsequent iterations can provide more coherent and precise control compared to using raw latent.

Furthermore, the numerical manipulation of latent vectors allows for greater creative freedom. Techniques include combining vectors to blend styles, expanding or upscaling the latent vector to enhance image resolution, or visualizing the latent process for improved interpretability. Additionally, it is possible to extract features from various images and integrate them into a single output image for better control.

In a blog post titled "Thought Vector," [21] Gabriel Goh discusses how thought vectors might represent a sparse and simplified data structure, where each vector is composed of several significant "atoms." Using face generation as an example, he demonstrates how exploring thought vectors can provide insights into what neural networks are processing and help debug their outputs. This method suggests a promising research direction for better understanding representational learning in neural networks.

However, it is crucial to recognize that latent vectors are not inherently interpretable by humans, posing challenges for direct manipulation using this technique.

2.1.3. Workflow III: Human-in-the-Loop Model

Human-in-the-Loop Model (Figure 2.3) marks a significant departure from the previous workflows by incorporating human feedback directly into the loop, emphasizing the dynamic interplay between AI and users within the creative process.

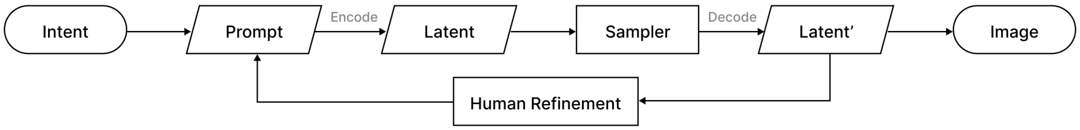


Figure 2.3: By incorporating human feedback into the loop, users can refine prompts based on their responses to the model, improving the resulting outputs.

The integration of human feedback is designed to correct and guide the AI's outputs, which may not always align with the artist's intentions. This aspect highlights the non-linear and iterative nature of artistic intent, which can evolve and adapt during the creative process.

Including humans in the loop allows artists to fine-tune the conditioning partially rather than changing the prompt at the beginning of the workflow.

By synchronizing the initial vision with the generated outcomes, this method enables artists to continually refine their prompts or results, ensuring alignment with their evolving creative goals.

A common practice involves artists importing AI-generated images into software such as Photoshop for further refinement or adjustments. These images can be manually edited or improved using inpainting models to target specific areas. For example, concept artists might initially use Midjourney to create preliminary concept art and later enhance the composition or integrate additional elements using Midjourney's inpainting tools or Photoshop's Generative Fill. Alternatively, artists may use the initial outputs as a foundation for further artistic development [40].

2.1.4. Workflow IV: Human-AI Parallel Model

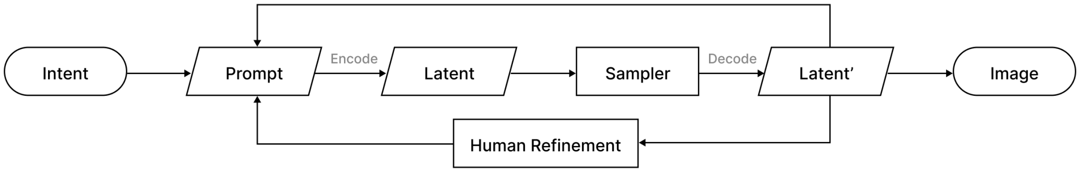


Figure 2.4: A dual feedback mechanism enables parallel processing of the AI's raw output and human-refined output to asynchronously update inputs for more nuanced modifications.

In the Human-AI Parallel Model (Figure 2.4), the Human-AI Parallel Model introduces a dual feedback mechanism that allows for parallel processing of the AI's raw output and human-refined output. This model offers artists a more nuanced approach to collaboration, where they can exert discrete control over separate feedback loops, thereby deepening their creative input and ownership over the resulting artwork. The interaction between the AI-generated results and human-modified outcomes forms a dynamic interplay, fostering a co-creative partnership that retains the artist's imprint beyond traditional models.

Sougwen Chung's projects, F.R.A.N. - Flora Rearing Agricultural Network[11] and Drawing Operations[10], exemplify this model. They showcase a harmonious interaction between artist and machine, with a robotic arm responding to and complementing Chung's gestures in a shared performance, reflecting the model's potential across various creative disciplines.

2.1.5. Workflow V: Visual Reference Model

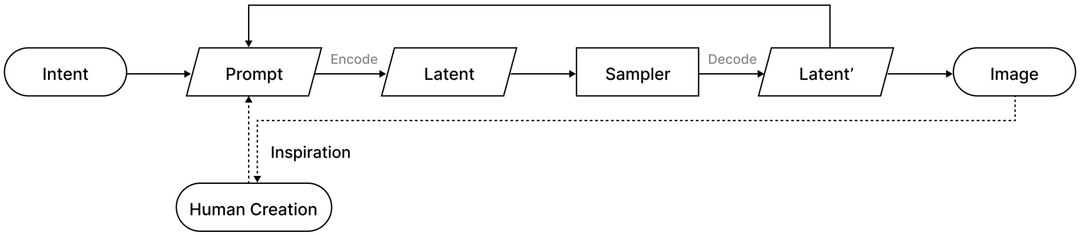


Figure 2.5: AI serves as a source of inspiration. Artists provide feedback on the input to refine the reference further.

The Visual Reference Model (Figure 2.5) utilizes the image generation model as a non-intrusive support system. In this setup, AI does not interfere with the artist's creative process; rather, it serves as a source of inspiration. Artists use outputs from the model as visual references, providing feedback and adjusting input prompts to refine these references further.

In the realms of art and design, a similar technique known as a moodboard [8] is an essential tool. It is a compilation of visual materials that encapsulate a specific theme, style, or concept, used by designers, illustrators, photographers, filmmakers, and other creatives to convey their vision for a project. Moodboards serve as a potent foundational element in any creative endeavor, offering a glimpse into the project's essence before the final pieces are made.

The intriguing aspect of this workflow is the potential for interaction between the moodboard and the artist's work. Unlike traditional moodboards, which are static, this interactive system would respond to the artist's ongoing work, dynamically adapting and providing enhanced creative possibilities. This responsive moodboard could revolutionize how creatives conceptualize and refine their ideas, making the creative process even more dynamic and fluid.

Chapter 3

Latent Auto-recursive Composition Engine (LACE)

The Latent Auto-recursive Composition Engine (LACE) was designed to bolster artistic control and enhance transparency in human-AI collaborations during the generative art creation process. LACE effectively tackles the typically opaque diffusion model sampling that is prevalent in AI-assisted image generation. By enabling the visualization of the denoising process, it empowers artists to directly influence the sampling trajectory. Furthermore, LACE integrates seamlessly with the Stable Diffusion ecosystem and Photoshop through ControlNet, allowing for precise manipulation of latent image features. This integration bridges the gap between complex numerical data and intuitive artistic editing. LACE's innovative approach not only enhances the artist's command over the generative model but also improves the clarity and authenticity of the creative process, thereby enriching the overall artistic experience. LACE's objective is to establish itself as a leading AI-based creative tool, offering artists a deeper sense of ownership and a more rewarding creative journey.

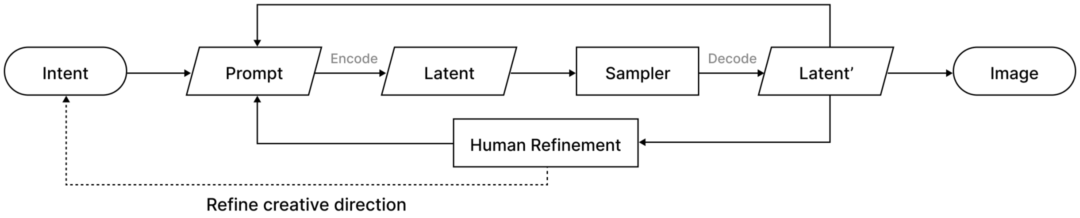


Figure 3.1: LACE features dual iterative loops, facilitating both AI-driven and human-influenced modifications in the creative process, ensuring alignment with the artist's evolving vision.

Section 3.1

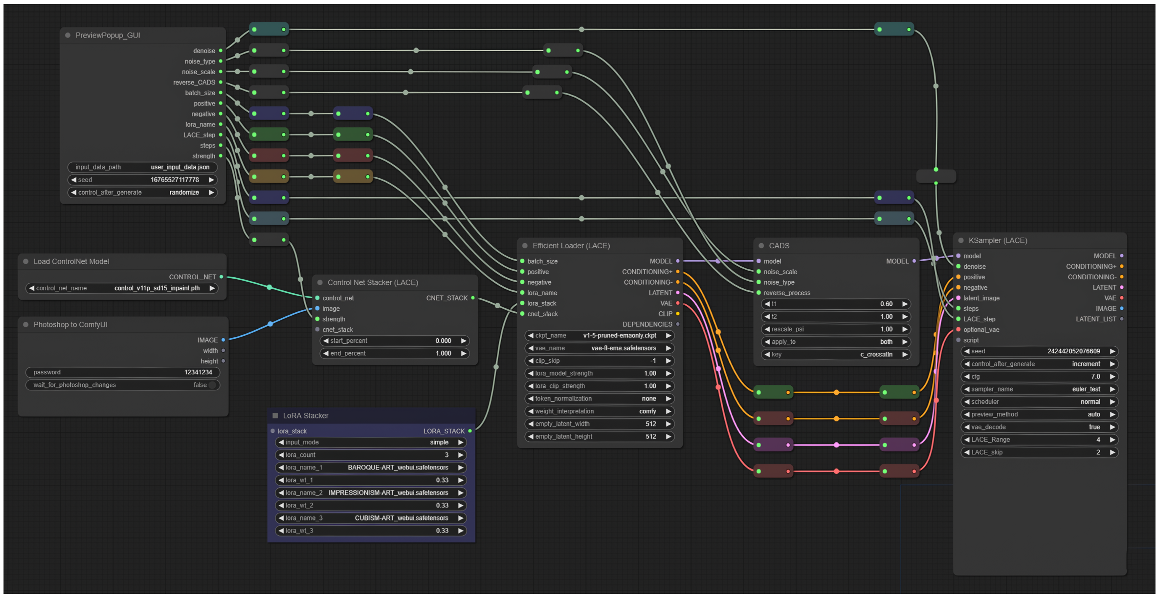
Architecture and Functionality

The LACE workflow, rooted in the Human-AI Parallel Model (workflow 2.1.4), provides artists with the capability to continuously update and refine their creative intent. Moving beyond the limitations of generative AI tools with static intent, LACE features dual feedback loops that enable users to adapt the AI pipeline in response to the evolution of their creative direction and to refine the generated results within Photoshop.

3.1.1. Engine Components and Workflow

LACE operates through two distinct iterative loops (Figure 3.1). The first loop utilizes Stable Diffusion v1.5 as the generative model, functioning without human intervention for initial image generation. The second loop introduces a human-in-the-loop mechanism, allowing artists to directly influence and modify the output by editing the image in Photoshop, concurrently with AI generation.

This dynamic interplay fosters a vibrant creative environment, producing artwork that is not only reactive and adaptable but also continuously evolves to align with the artist's changing vision while navigating through latent space.



input_data_path user_input_data.json
seed 16765527117778
control_after_generate randomize

control_net_name control_v11p_sd15_inpaint.pth

password 12341234
wait_for_photoshop_changes false

start_percent 0.000
end_percent 1.000

input_mode simple
lora_count 3
lora_name_1 BAROQUE-ART_webui.safetensors
lora_wt_1 0.33
lora_name_2 IMPRESSIONISM-ART_webui.safetensors
lora_wt_2 0.33
lora_name_3 CUBISM-ART_webui.safetensors
lora_wt_3 0.33

ckpt_name v1-5-pruned-emaonly.ckpt
vae_name vae-ft-ema.safetensors
clip_skip -1
lora_model_strength 1.00
lora_clip_strength 1.00
token_normalization none
weight_interpretation comfy
empty_latent_width 512
empty_latent_height 512

t1 0.60
t2 1.00
rescale_psi 1.00
apply_to both
key c_crossattn

seed 242442052076609
control_after_generate increment
cfg 7.0
sampler_name euler_test
scheduler normal
preview_method auto
vae_decode true
LACE_Range 4
LACE_skip 2

Figure 3.2: The render pipeline of LACE in ComfyUI

In constructing the LACE workflow (Figure 3.2) within ComfyUI[13], a modular approach was adopted, capitalizing on the node-based architecture's versatility. Central to the pipeline's architecture are the nodes from efficiency-nodes-comfyui[38], forming the backbone of the system.

To address the issue of homogeneity in Stable Diffusion's output, the CADS mechanism—inspired by the principles of Condition-Annealed Sampling as illustrated by Sadat et al.(2024) [37]—was assimilated to enhance the variety and influence within the generative procedure.

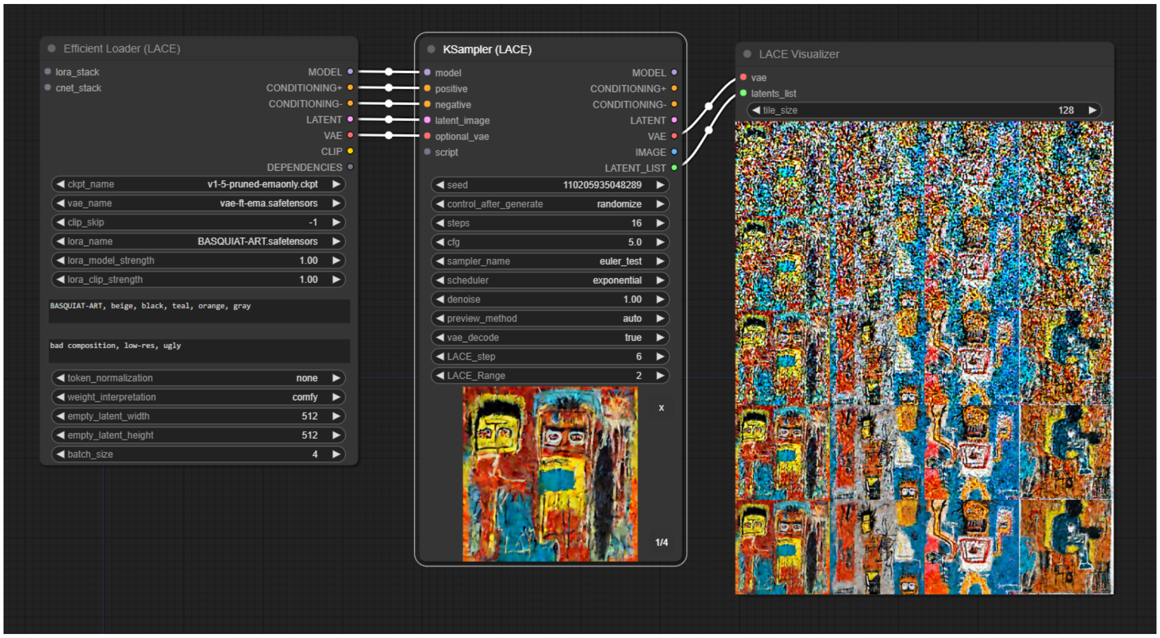
For the image prompt, the "comfyui-photoshop" node [29] is adapted and refined, which acts as an observer to Photoshop's localhost, leveraging Photoshop's built-in features. The image prompt is subsequently relayed to ControlNet as the latent input for the Efficient Loader.

To manage the interaction between the nodes, the "PreviewPopup-GUI" node is employed, which acts as a traffic director, guiding data along the pipeline, and providing artists with the ability to modify pipeline settings directly within Photoshop.

Thereby reducing the need to toggle between the ComfyUI backend and Photoshop.

The "Efficient Loader (LACE)" node, is adeptly integrated with the Stable Diffusion v1.5 model [36]. This pivotal node kickstarts the generative sequence by loading the essential models and prompts. Additionally, a refined VAE, sd-vae-ft-ema[3], developed by Stability.AI is adapted, enhancing the image fidelity of the creative output.

The "KSampler (LACE)" node is tailored to punctuate the generative process at designated intervals through the "LACE step" feature, enabling the capture of latent representations at different denoising stages. This functionality offers insight into the internal dynamics of the model, affording artists a transparent view of the image synthesis evolution. Moreover, it permits users to preserve the latent state at a specified timestep, which can then be re-injected into future sampling iterations for enhanced directional influence.



ckpt_name v1-5-pruned-emaonly.ckpt
vae_name vae-ft-ema.safetensors
clip_skip -1
lora_name BASQUIAT-ART.safetensors
lora_model_strength 1.00
lora_clip_strength 1.00

```
BASQUIAT-ART, beige, black, teal, orange, gray
```

```
bad composition, low-res, ugly
```

token_normalization none
weight_interpretation comfy
empty_latent_width 512
empty_latent_height 512
batch_size 4

seed 110205935048289
control_after_generate randomize
steps 16
cfg 5.0
sampler_name euler_test
scheduler exponential
denoise 1.00
preview_method auto
vae_decode true
LACE_step 6
LACE_Range 2

tile_size 128

Figure 3.3: The visualization of the denoising process in LACE

Furthermore, the "LACE step" (Figure 3.3) doubles as a preview mechanism that accelerates the inference speed. In the latter stages of denoising, the compositional elements of the image remain largely unaltered; it is the finer details that are honed. Thus, early previews can be leveraged for quicker iterative feedback without compromising the structural integrity of the final image.

The "LoRA Stacker" node introduces learned stylistic modifications to the raw Stable Diffusion v1.5 model, functioning as an artistic palette within the generative pipeline. This stacker facilitates the infusion of unique artistic styles, which artists can finely calibrate. Analogous to blending pigments in painting, this node allows for the precise mixing of different weights from the LoRA models, enabling the artist to craft a customized style in the generated images.

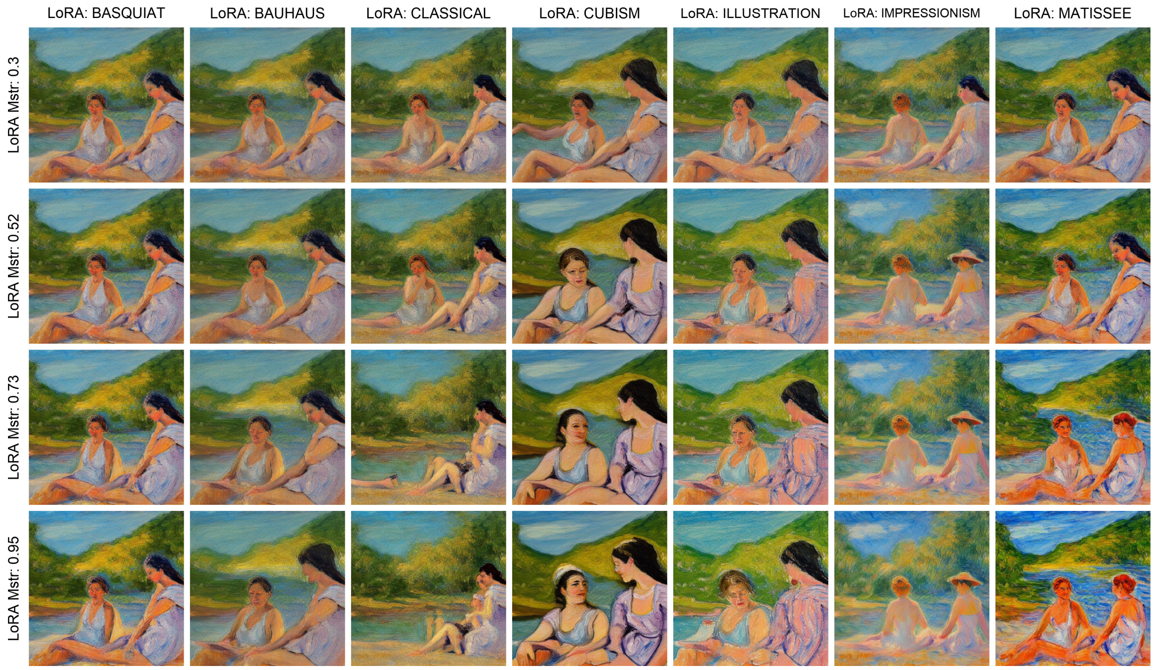


Figure 3.4: Exploration of artistic styles, composition, and texture using mixed weights from different LoRA models applied to a base model (LoRA: impressionism). Prompt used: "a image of a painting of two women sitting on a beach".

3.1.2. User Interface and Control Mechanisms

To enhance the user experience, an external GUI has been developed (Figure 3.5), utilizing Tkinter, the standard Python interface to the Tcl/Tk GUI toolkit [32], designed to integrate seamlessly with Photoshop. This GUI actively monitors and captures parameter updates, subsequently saving them to a JSON file. This file acts as the central communication hub between Photoshop and the AI pipeline interfaced through the "PreviewPopup-GUI" node.

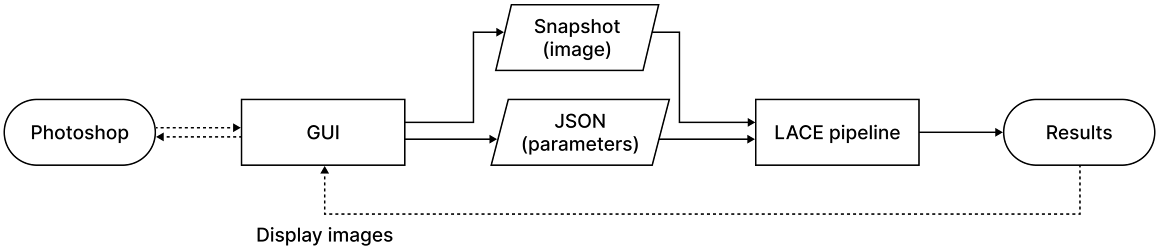
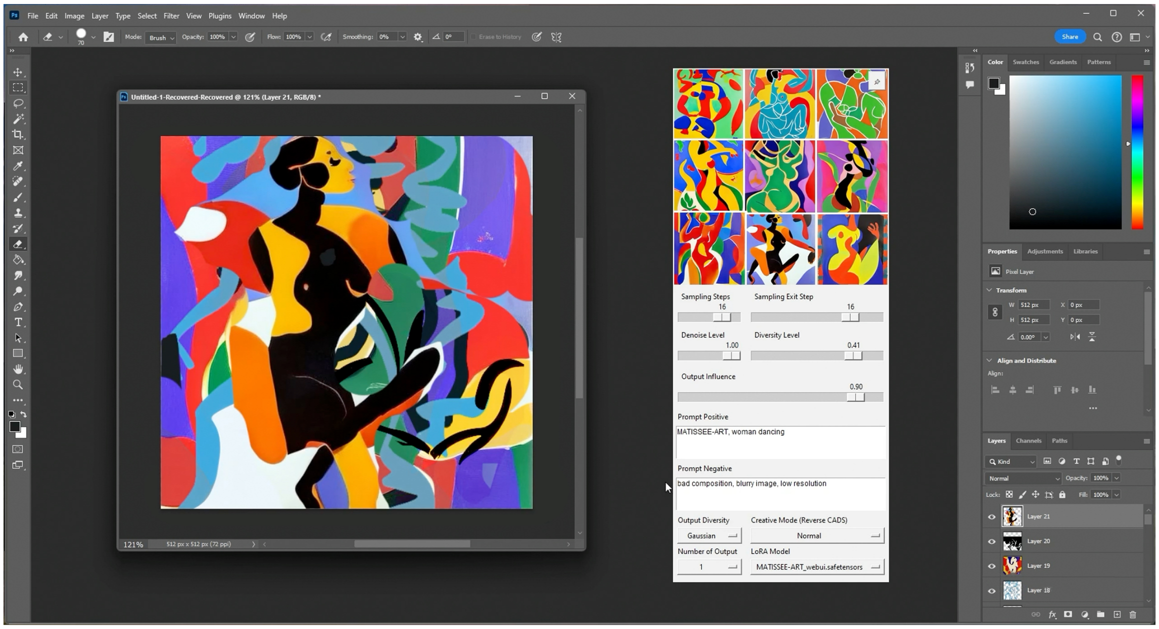


Figure 3.5: The Structure and Information Pathway of the LACE GUI

The GUI (Figure 3.6) is equipped with various input fields allowing for the adjustment of parameters such as text prompts, batch size, LACE step, denoise level, CADS[4], and the ControlNet strength, as well as the selection of the LoRA model. These inputs are later translated into user-friendly descriptions, making the interface accessible to users without an extensive background in AI technology.

The overall LACE architecture culminates in a system that not only reflects but also enhances the artist's creative intent. It is a platform where the generated image is not an end point but a dynamic element in a continuous dialogue between the artist and the AI. This dialogue is the heart of the LACE workflow—a system that is both a canvas and a crucible for the synthesis of new art.


File Edit Image Layer Type Select Filter View Plugins Window Help
Untitled-1-Recovered-Recovered @ 121% (Layer 21, RGB/8) *
121%
512 px x 512 px (72 ppi)
Color Swatches Gradients Patterns
Properties Adjustments Libraries
Pixel Layer
Transform
Align and Distribute
Layers Channels Paths
Layer 21
Layer 20
Layer 19
Layer 18
Share

Sampling Steps 16

Sampling Exit Step 16

Denoise Level 1.00

Diversity Level 0.41

Output Influence 0.90

Prompt Positive: MATISSEE-ART, woman dancing

Prompt Negative: bad composition, blurry image, low resolution

Output Diversity: Gaussian

Creative Mode (Reverse CADS): Normal

Number of Output: 1

LoRA Model: MATISSEE-ART_webui.safetensors

Figure 3.6: GUI of LACE in Adobe Photoshop

Section 3.2

Enhancing Interpretability and Artistic Engagement

3.2.1. Addressing the AI "Black Box"

To avoid the complexities of direct latent manipulation, which may not provide a clear understanding of how inputs are transformed into outputs, we employ an image-to-image approach. In this method, latents are initially decoded into images, which are then edited in Photoshop and subsequently re-encoded into latents via ControlNet. This approach significantly enhances the model's explainability by utilizing Photoshop's user-friendly image manipulation tools, allowing for more precise and subtle adjustments than those achievable with text prompts alone, thereby more effectively guiding the sampling process.

A fundamental objective is to maintain a precise and predictable image-to-image feedback loop. This is vital to ensure that modifications made in Photoshop do

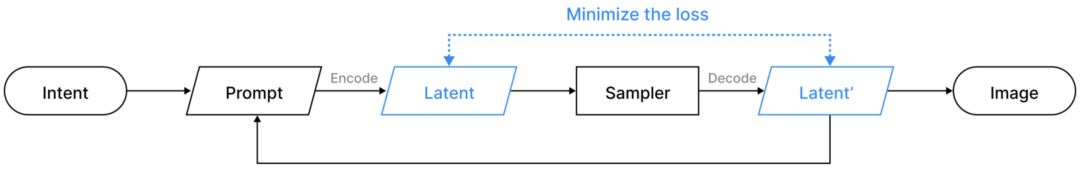


Figure 3.7: This figure illustrates the minimal transformation from input to output within the LACE system, ensuring that users can easily understand how their edits affect the final image in photo manipulation.

not introduce unpredictability into the system, as depicted in Figure 3.7. To verify the effectiveness of LACE, it is crucial that the transformation from input to output remains minimal, allowing users to clearly discern the relationship between their edits and the resultant image in photo manipulation.

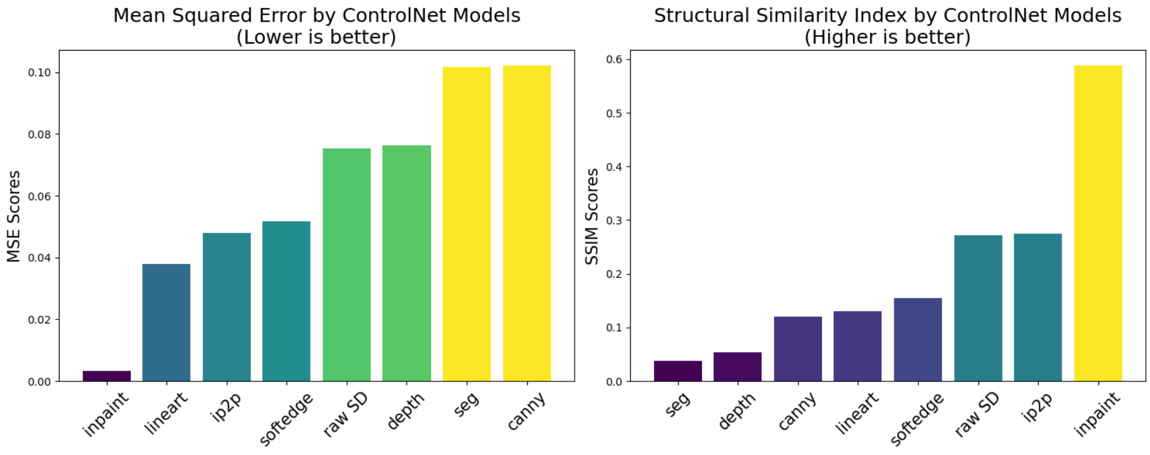


Figure 3.8: Comparison of Stable diffusion's native VAE and ControlNet-enhanced VAE in reproducing input images

An experimental evaluation was conducted to compare the performance of the native VAE encoder in Stable Diffusion 1.5 with a ControlNet-enhanced version. This assessment focused on their ability to accurately reproduce an input image without the aid of a text prompt. The procedure involved encoding an image into a latent form, sampling a new latent based on the initial encoding, and decoding it back to an image to observe the fidelity of reproduction.

As a result (Figure 3.8), ControlNet significantly surpasses the native encoder in terms of SME (square mean error in RGB pixels) and SSIM (structural similarity index), supporting the use of LACE for creating interactive feedback loops for iterative artistic refinement.

To further analyze how the loss between the original and sampled latent evolves, an experiment was conducted on how a reference image changes over 20 iterations by re-encoding the current output as the next input latent and continuing the sampling process under consistent parameters.

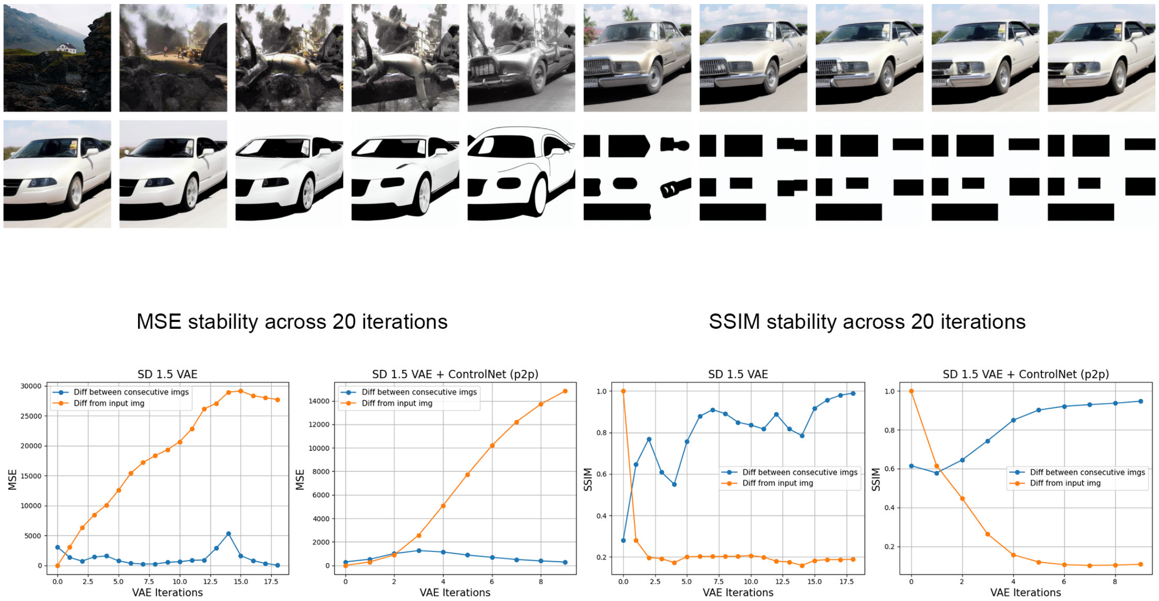


Figure 3.9: Tracking MSE and SSIM across 20 iterations of Stable Diffusion's VAE, regenerating an input image without additional prompts. Each iteration transforms a mountainous landscape into abstract geometric forms..

These experiments (Figure 3.9) confirmed that MSE and SSIM values for ControlNet are more stable across consecutive iterations. The gradual loss curve suggests that outputs from the image-to-image mode facilitated by ControlNet are more predictable, further validating the use of an image-based input approach in LACE facilitated through Photoshop.

3.2.2. Facilitating Artistic Decision-Making

LACE, developed as a human-AI parallel system (see 2.1.4), reintroduces the edited latent as a new prompt for subsequent sampling. This system can operate in two modes:

Mode 1: Text-Prompt-Based Method - Initially, users may use positive and negative prompts to navigate the latent space until they achieve an image close to their artistic vision. Subsequently, they can import this image into Photoshop for detailed editing, such as adjusting color tones or compositions. This method significantly enhances creative freedom by combining Photoshop's powerful editing capabilities with AI-driven rendering, helping users merge various elements seamlessly.

Mode 2: Sketch-Based Method - Alternatively, users can start with a sketch or image collage in Photoshop as the initial input instead of a text prompt. This approach reduces the unpredictability of AI outputs by clearly defining the artistic direction through visual means before any textual descriptions are introduced, which might lead the process astray.

Both methods are designed to incorporate human judgment into the generative process, addressing the opacity of the AI 'black box' by ensuring that human creativity plays a central role in the creation process.

3.2.3. Enhancing Creativity with Condition-Annealed Sampling

For art and design, it is helpful that AI supports brainstorming with diversified outputs. It can provide visual suggestions and accelerate the creative process. However, conditional diffusion models suffer from low output diversity when sampled with high classifier-free guidance [23] scales for optimal quality, or when trained on small datasets. This results in generated samples that look very similar despite using different random seeds. Thus, in LACE, a function is implemented for more diverse results empowered by Condition-Annealed Diffusion Sampler (CADS) based on Seyedmorteza Sadat et al. paper in 2024 [37].

A node is built in ComfyUI to address this problem by gradually annealing the conditioning signal during the sampling process. It adds noise (Gaussian, Uniform, Exponential) to the conditioning embedding, starting with a high noise level and decreasing it to zero by the end of sampling, or the opposite to create more extreme results. This allows more diversity in the early sampling steps while still respecting the conditioning at the end.

$$\hat{\mathbf{y}} = \sqrt{\gamma(t)}\mathbf{y} + s\sqrt{1 - \gamma(t)}\mathbf{n} \tag{3.1}$$

In the given equation, $\hat{\mathbf{y}}$ denotes the noisy condition vector, $\mathbf{y}$ represents the original condition vector, and s specifies the initial noise scale. The term $\mathbf{n}$ is Gaussian noise sampled from a standard normal distribution $\mathcal{N}(0, I)$, which can alternatively be replaced with uniform or exponential distributions. The function $\gamma(t)$ acts as an annealing schedule, decreasing from 1 to 0 as t transitions from 1 to 0 throughout the reverse sampling process.

The annealing schedule $\gamma(t)$ is defined as a piece-wise linear function:

$$\gamma(t) = \begin{cases} 1 & t \leq \tau_1, \\ \frac{\tau_2 - t}{\tau_2 - \tau_1} & \tau_1 < t < \tau_2, \\ 0 & t \geq \tau_2, \end{cases} \tag{3.2}$$

for user-defined thresholds $\tau_1, \tau_2 \in [0, 1]$. Since diffusion models operate backward in time from $t = 1$ to $t = 0$ during inference, the annealing function ensures high corruption of $\mathbf{y}$ at early steps and no corruption for $t \leq \tau_1$. The annealing function can also be reversed to corrupt $\mathbf{y}$ at the end, making more extreme results.

Adding noise to the condition changes the mean and standard deviation of the conditioning vector. In most experiments, the paper rescales the conditioning vector back toward its prior mean and standard deviation. Specifically, for a clean condition vector $\mathbf{y}$ with (scalar) mean and standard deviation μ_{in} and σ_{in}, we compute the final corrupted condition $\hat{\mathbf{y}}final$ according to

$$\hat{\mathbf{y}}_{rescaled} = \frac{\hat{\mathbf{y}} - \text{mean}(\hat{\mathbf{y}})}{\text{std}(\hat{\mathbf{y}})}\sigma_{in} + \mu_{in} \tag{3.3}$$

$$\hat{\mathbf{y}}_{final} = \psi\hat{\mathbf{y}}_{rescaled} + (1-\psi)\hat{\mathbf{y}}, \tag{3.4}$$

for a mixing factor $\psi \in [0, 1]$. This rescaling scheme prevents divergence, especially for high noise scales s, but slightly reduces the diversity of the outputs. Therefore, one can trade more stable sampling with more diverse generations by changing the mixing factor ψ.

Section 3.3
Experiment: Applying LACE to Increase the Sense of Agency in Artistic Creation

To assess the effectiveness of LACE, an experiment evaluates the impact of artist involvement in decision-making on the interpretability of AI models and the alignment of expected outcomes with actual results. The study aims to determine how these dynamics influence artists' sense of ownership and engagement over their creations and explore whether a diminished role in the creative process could devalue the artwork from the artists' perspective due to perceived minimal human input.

To unpack the interaction between artists and AI systems in creative workflows, the study employed Stable Diffusion to evaluate across three primary creative workflow patterns, as detailed in Section 2.1: Analysis of Human-AI Collaborative Workflow, which provides a framework for understanding artist engagement and influence on AI systems during the artistic creation.

3.3.1. Experimental Design

Objective: The experiment is designed to gauge the degree of agency that users experience when engaging with different AI-driven creative workflows. A focal point of the study is to determine the effectiveness of the "human-in-the-loop" approach in reinforcing the presence of a human touch within the generated art.

Hypothesis: The research hypothesizes that a "human-in-the-loop" workflow will notably increase the user's sense of agency and engagement in comparison to conventional text-to-image generation methods. Furthermore, it posits that the implementation of LACE within the Stable Diffusion model will make it more interpretable. As a result, users are expected to value their artwork more due to their increased involvement in the creation process.

Participant demographics: For participant recruitment, a total of 21 students from Dartmouth College with an interest in digital art will be selected.

Test Materials: The materials for testing (Table 3.1) encompass diverse facets of creative expression to evaluate user interaction with different prompt types:

- T1 (Representational): Participants are tasked to illustrate a scene depicting "a man reaching for a painting in an art gallery, accompanied by a dog sniffing another artwork on the floor."
- T2 (Non-Representational): Users are prompted to create "an abstract composition that embodies the dynamics and motion associated with joy."

- T3 (Design Challenge): The assignment involves designing “a pixel art game scene with a bustling cityscape featuring assorted architectural styles.”

These materials are intended to test participants’ responses to varying prompt complexities. Representational prompts, which are typically easier to visualize, allow for straightforward interpretation, while non-representational prompts demand higher creative input from users. Moreover, the design challenge is set to investigate whether the “human-in-the-loop” approach can imbue greater depth and nuance into the outcomes compared to traditional prompt-based engineering, especially as CLIP models often struggle to synthesize design concepts through text prompts alone.

Task Design: Participants will engage with three distinct workflows designed to test various methods of interaction with the AI model:

- **W1 (Single Sampler with Text Prompt)** 2.1.1: In this workflow, participants use a single text prompt to iteratively refine their artwork until it aligns with the given task prompt. This method employs a straightforward, end-to-end process where the same sampler is repeatedly used for image generation.

- **W2 (Multiple Samplers with Text Prompt)** 2.1.2: This approach expands on the first by utilizing multiple samplers in sequence to evolve the artwork. Here, each sampler’s input is derived from the output of the previous sampler, enhancing the interpretability of the process by visualizing the step-by-step development of the image.

- **W3 (LACE with Text and Image Prompt)** 2.1.4: This method integrates Photoshop editing and ControlNet to allow participants to directly manipulate the generated images. The edited images are then fed back into the system as image prompts along with text, creating a dynamic feedback loop. This setup is supported by a GUI that lets users adjust the parameters of the Stable Diffusion model to influence the generation process actively.

These workflows are crafted to compare traditional text-based prompting with more interactive and iterative approaches that might enhance user engagement and model transparency.

Experiment flow: Participants are randomly assigned one of the creative prompts (T1, T2, T3) and will explore all three workflows (W1, W2, W3) using the assigned prompt to ensure a comprehensive assessment across different methods. The order in which they engage with the workflows is randomized to prevent fatigue bias from affecting the results. At the start of each task, participants receive a brief training session on the prototype, accompanied by a demonstration to familiarize them with the process. They are informed that they may conclude the task at any point before a covert 15-minute time limit expires. Immediately after completing each workflow, participants are required to fill out a survey (Table 3.2) to capture their immediate responses and impressions. Upon finishing all tasks, a brief open-ended interview is conducted to gather additional qualitative feedback and provide insights into their experiences with each workflow configuration.

Table 3.1: Testing Materials

Type	**Art Prompt**	**Workflow**	**Duration**
T1: Representational	a man reaching for a painting in an art gallery, accompanied by a dog sniffing another artwork on the floor.	W1, W2, W3 (Random Order)	15 mins each workflow
T2: Non-Representational	an abstract composition that embodies the dynamics and motion associated with joy.	W1, W2, W3 (Random Order)	15 mins each workflow
T3: Design Challenge	a pixel art game scene with a bustling cityscape featuring assorted architectural styles.	W1, W2, W3 (Random Order)	15 mins each workflow

Table 3.2: LACE User Testing Survey Questions

Section	**Questions**
Background Information	- Major - Years of Photoshop experience - Experience in art or digital art - Computer science background - Experience with generative AI products
Survey After Each Workflow	- How many minutes do you think Task 1 took to complete? - How satisfied are you with the final result? - How much do you feel a sense of ownership over the final result? - To what extent does the final output diverge from your initial expectations? - How would you rate the explainability of the tool (workflow) provided? - How would you rate the usability of the tool (workflow) provided to complete the task? - Would you consider the final output to be 'Art'?
Final Review	- Which workflow do you enjoy the most? - Which workflow do you enjoy the least? - Additional comments

Chapter 4

Results and Discussion

To explore how LACE and various other workflows affect the creative process, an experiment was conducted involving 21 participants. Each session lasted approximately one hour, and the entire testing spanned 10 days, conducted in the ILIXR Lab (ECSC Room 102). In the following section, we will discuss some of the key findings and observations.

Section 4.1 Demographics

The demographic composition of the study's participants (Figure 4.1), all of whom are Dartmouth students, indicates a possible institutional bias inherent in the sample group. This selection comes from a diverse array of academic backgrounds, with the largest proportions being from computer science with a focus on digital art (47.6%) and a significant representation from the broader computer science discipline (28.6%). Additionally, the sample included students from varied fields such as the Japanese language, biomedical engineering, and geology, albeit in smaller percentages. In terms of proficiency with digital tools, the participants' experience with Adobe Photoshop spread across a spectrum: the majority (45%) had less than one year of experience,

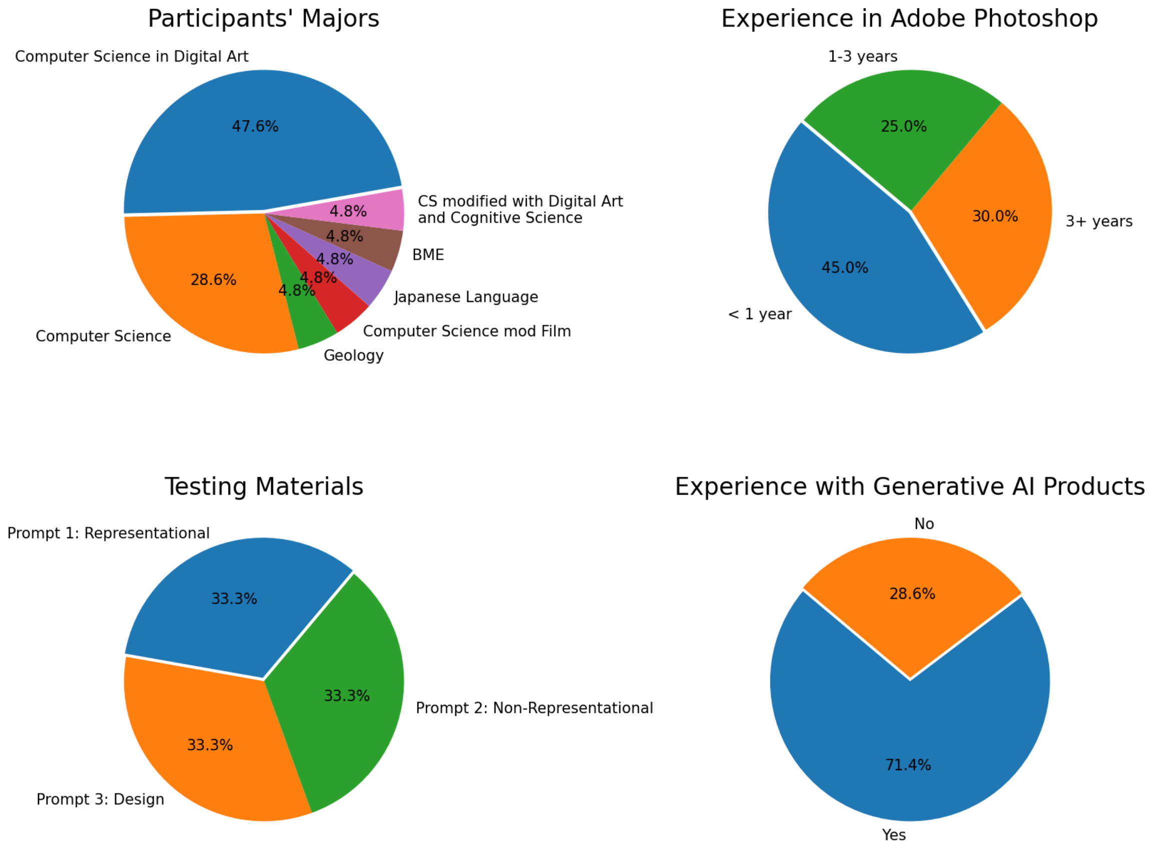


Figure 4.1: Overview of the participants

a quarter had navigated the software for 1-3 years, and the remainder (30%) were more seasoned users with over three years of experience. The experimental prompts were meticulously distributed, with each of the three categories—representational, non-representational, and design—receiving an equal share of focus. Furthermore, a significant majority (71.4%) of the participants reported having prior experience with generative AI products, an aspect that could influence their engagement with the testing materials and the resulting data. It's also important to disclose that some of the participants have personal ties to the researcher, which could conceivably introduce a subjective skew to the data. For a detailed breakdown of these demographics and their potential implications on the research outcomes.

Section 4.2

Scoring and Key Findings

4.2.1. Overview in Qualitative Survey

Figure 4.2 presents a comparative evaluation of three workflows—W1 (Single Sampler with Text Prompt), W2 (Multiple Samplers with Text Prompt), and W3 (LACE with Text and Image Prompt)—across six metrics. The results demonstrate that W3 (LACE) outperforms the other workflows, particularly in the categories of **'Consider as Art,' 'Model Explainability,' 'Usability,' 'Ownership,'** and **'Result matches expectations.'**

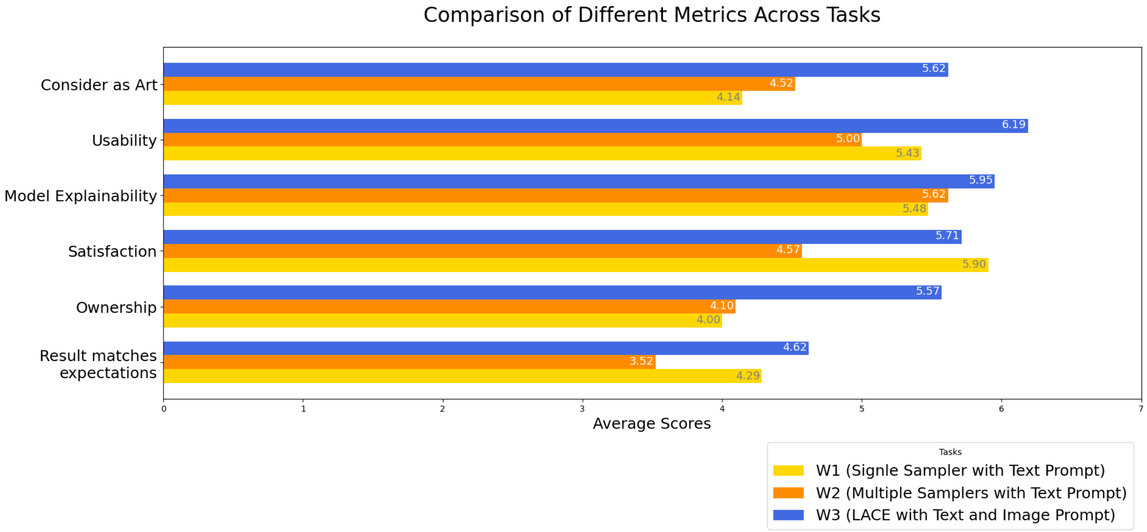


Figure 4.2: Overview of all metrics.

Interestingly, while W3 (LACE) is generally favored for its artistic association and usability, W1, which employs the most common method of iterative prompt engineering in text-to-image models, scores the highest in satisfaction with the final output. This finding suggests that, with the same quality of output achieved by using the same model, satisfaction might not always correlate with ownership, explainability, or the degree of control.

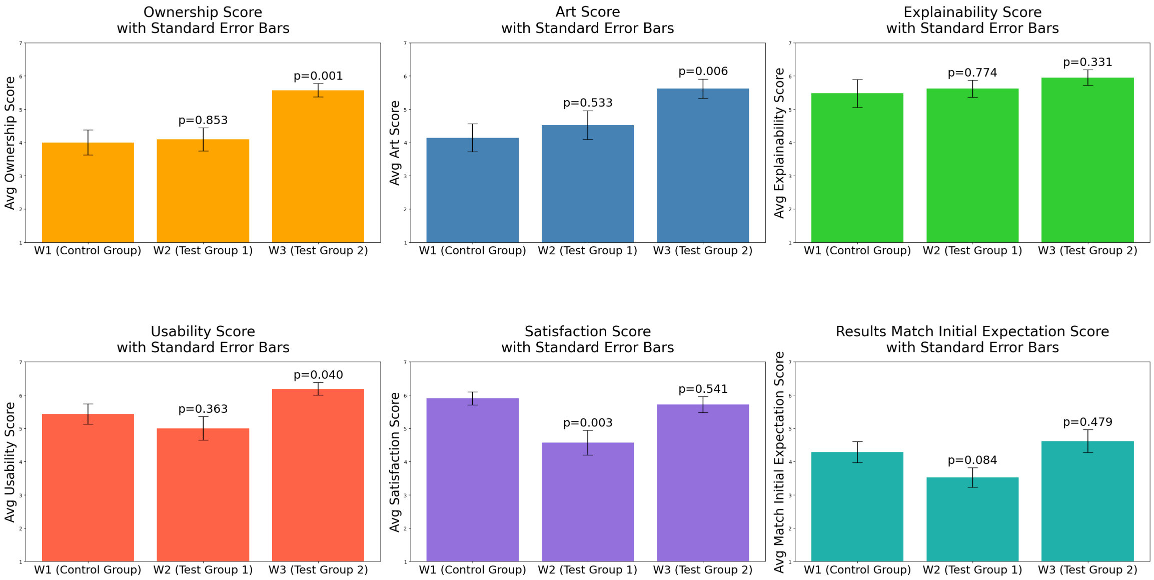


Figure 4.3: T-test and Standard Error across all metrics. The p-values are derived from T-tests comparing W1 with W2 and W1 with W3.

Surprisingly, W2, which aims to visualize the denoising process in diffusion models, has lower scores than W1 in usability, satisfaction, and result matching expectations. This could be attributed to the fact that participants might not have sufficient knowledge of how diffusion models and the denoising process work, potentially impacting their perception and understanding of the workflow.

To determine if W3 (LACE) significantly outperforms other workflows, a t-test analysis was conducted on the six metrics, as shown in Figure 4.3. The results reveal that only **'Usability'** and **'Ownership'** are significantly higher in W3 compared to other workflows, with p-values of 0.04 and 0.01, respectively. This finding suggests that despite the human-in-the-loop approach employed in W3, the explainability of the AI model remains insufficient. The lack of significant improvement in model explainability might potentially explain why participants did not consider their final work as art and felt that it did not match their initial expectations. This highlights the importance of enhancing the interpretability and transparency of AI models in creative workflows to foster a stronger sense of artistic ownership and satisfaction among users.

4.2.2. Engagement and Agency

To assess participant engagement and sense of agency in the creative process enabled by generative AI, the study—inspired by James W. Moore and Sukhvinder S. Obhi [28]—required participants to self-report the time spent on each task immediately following its completion. Simultaneously, researchers recorded the actual time spent on the tasks for comparison. This method not only illuminated participants' engagement levels but also allowed for the exploration of discrepancies between perceived and actual time investments, potentially uncovering cognitive biases or differences in task absorption.

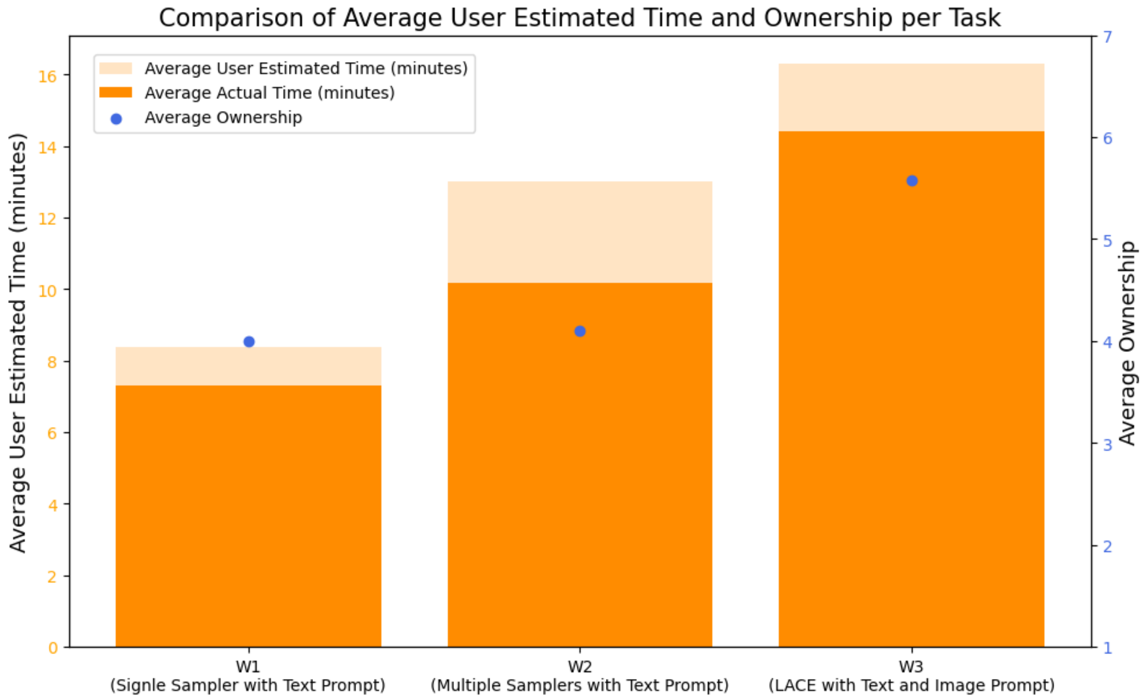


Figure 4.4: Both the time and the ownership scores increased progressively from W1 to W3

From Figure 4.4, we observe that participants consistently overestimated the time they spent on tasks across all three workflows. The discrepancy between perceived and actual time spent was greatest in W2 (Multiple Samplers with Text Prompt) and least in W1 (Single Sampler with Text Prompt). This trend suggests a potential correlation between the complexity or usability of the workflow and participants' time perception. Testers may overestimate their engagement due to the cognitive load required to learn or address the challenges they encounter.

However, a greater discrepancy in time estimation does not necessarily translate to a higher sense of agency. Despite W2 showing significant differences in both estimated and actual time, it did not lead to increased feelings of ownership.

On the other hand, W3, which employed the LACE workflow incorporating human-in-the-loop and iterative reflection on the creative intent, recorded the highest average ownership scores. This pattern suggests that integrating human-in-the-loop and visually oriented LACE workflows might enhance participants' sense of ownership and their connection to the created content. This could be attributed to the workflow's interactive nature and the added dimension of intentional binding, which intensifies personal engagement and resonance with the creative process.

4.2.3. Task Complexity and User Effort

An additional aspect of engagement is analyzed by examining the relationship between the time spent on tasks and participant satisfaction, categorized into three levels: low (1-2), medium (3-5), and high (6-7).

Figure 4.5 reveals an interesting pattern where the time spent on each workflow progressively increases in the high satisfaction group. For the medium satisfaction group, the time spent is fairly consistent across workflows without notable differences. Interestingly, the low satisfaction group shows no entries for low satisfaction in W1, but a significant increase in both estimated and actual time spent is observed in the other workflows.

This pattern suggests that tasks perceived as less satisfying may seem more time-consuming to participants, potentially indicating a higher cognitive load or lower enjoyment, which could in turn affect their perception of time.

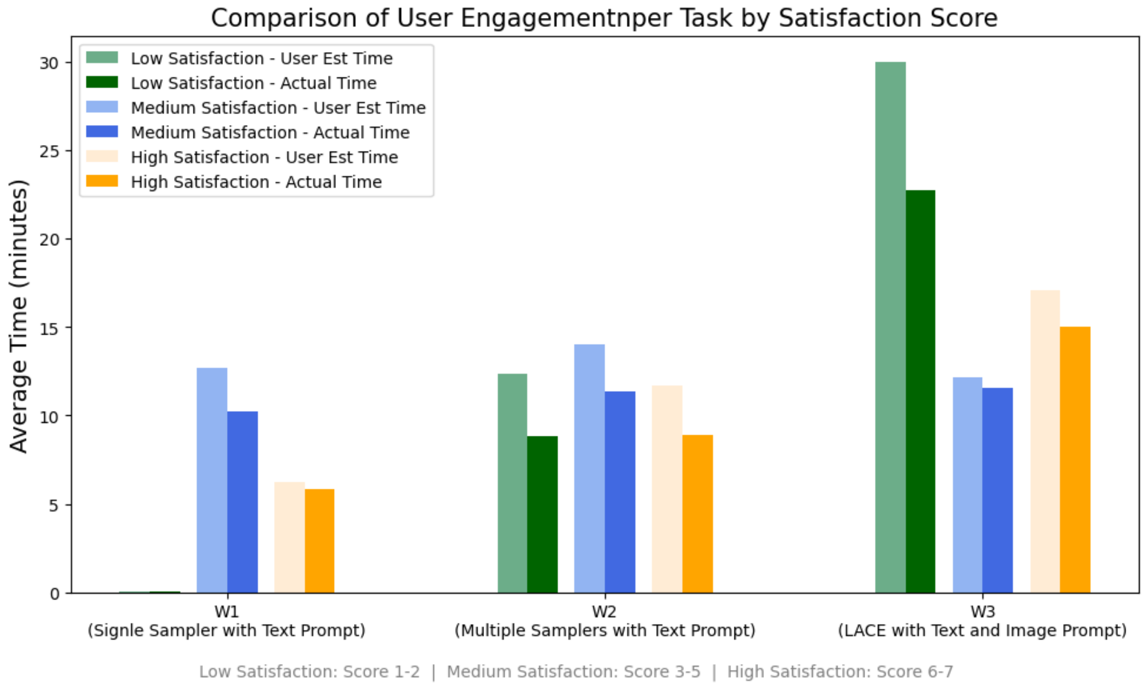


Figure 4.5: Tasks that are perceived as less satisfying may feel more time-consuming to participants, potentially pointing to a higher cognitive load or reduced enjoyment.

4.2.4. Comparison of Times by Task and Prompt

Another perspective on engagement considers the type of art prompt given to participants. Figure 4.6 shows an upward trend in both estimated and actual time from W1 to W3 for T1 and T3, with estimated times consistently surpassing the actual times spent on tasks. This suggests an overestimation of effort if participants have a clear vision of what to work on.

Conversely, T2 revealed an interesting anomaly where participants' estimated times were lower than the actual times in both W1 and W3. This could suggest that when faced with a less concrete, more abstract task requiring imagination, as in T2, participants may underestimate the effort required or engage with the task more superficially, resulting in an actual time investment that exceeds their estimations.

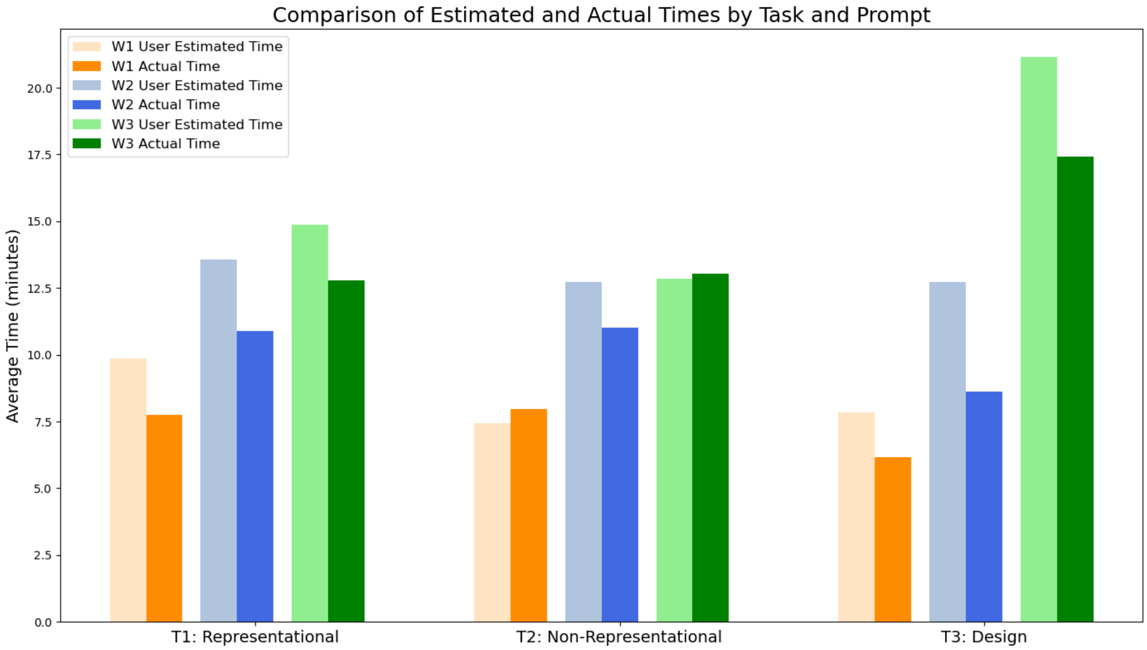


Figure 4.6: This graph displays trends in estimated and actual times by task and prompt type, highlighting differences in participant engagement and perception.

This trend might imply that participants, particularly those with less background in traditional art, find it challenging to connect with the abstract prompts, potentially leading to reduced effort or engagement.

Section 4.3 Discussion

Through open interviews and research observations, a comprehensive understanding of user interactions with various workflows was established. The quality of the final images proved to be a pivotal factor in participant satisfaction. High-quality outputs substantially improved the user experience, whereas lower-quality results, particularly in intricate tasks like pixel art, decreased satisfaction. It was noted that the pursuit of diversity within the AI models occasionally compromised quality, emphasizing the necessity for a strategic balance between diversity and excellence.

W3 (LACE) received commendation for its practical utility, fostering creative exploration, and enhancing users' sense of ownership over their creations. Nonetheless, difficulties in thumbnail assessment highlighted the need for improvements in user interface design. Users' proficiency levels varied, and discrepancies between user intent in prompts and AI interpretations often led to frustration, suggesting a requirement for more intuitive interface designs, especially to manage the rapid and unpredictable changes observed in the Photoshop workflow. Those proficient in Photoshop could effectively leverage its features in photo manipulation, indicating that compatibility with users' existing skill sets is crucial for tool adoption.

W2 received mixed feedback regarding its customization level. Participants with an AI background appreciated the freedom to personalize their art, whereas others found the complex controls daunting, which prolonged the learning curve and imposed time constraints. This workflow was particularly valued for its capacity to incorporate specific user elements into AI-generated outputs, deepening the personal connection to the work.

Across all workflows, the unpredictability of AI interpretation of textual prompts posed a consistent challenge, especially in the T2 (non-representational) task with abstract concepts such as "joy." This gap often left participants dependent on AI guidance, highlighting a significant area for improvement in human-AI collaboration. Despite the integration of human-in-the-loop systems like W3, aligning AI-generated images with users' conceptual visions remains a challenging goal, underscoring the inherent unpredictability of the current state of text-to-image AI models.

Section 4.4

Limitations

The demographic limitations of this study, centered solely around Dartmouth College students with backgrounds in computer science or digital art, may inhibit the extrapolation of its findings to a wider audience.

The absence of traditional art expertise among participants might have colored their judgments when addressing queries such as “Would you consider the final output to be ‘Art’?” Those versed in art could provide insights that reflect a broader conceptual grasp of artistic outputs. Moreover, the majority of participants’ limited experience with Photoshop, predominantly under a year, could have hampered their capacity to fully engage with W3, thereby affecting the study’s results.

Within W3, it was observed that participants rarely engaged with the LACE as an auxiliary visual tool, similar to the approach in workflow V (Section 2.1.5). Instead, they predominantly adhered to a method centered around text prompts. This behavior indicates that the interface was often perceived merely as a sophisticated text-to-image conversion tool rather than a fully-fledged visual assistant. Such a trend highlights the importance of conducting further research with W3 across diverse settings and with participants from a wide array of backgrounds. This would help reduce potential biases and strengthen the trustworthiness of the study’s conclusions.

Additionally, the researcher observed that participants generally had minimal experience with the intricacies of prompt engineering in text-to-image models. The underuse of certain “magic words” that could notably refine the output quality suggests a gap in user expertise. To address this, the onboarding process was adapted to encourage participants to concentrate on artistic design rather than output quality, a modification that may have influenced both test outcomes and participant behavior.

This evolution in the research methodology reflects the dynamic nature of user interaction studies and highlights the importance of adaptability in experimental design to better support user needs and research objectives.

Chapter 5

Conclusion

Section 5.1

Conclusion

The integration of a human-in-the-loop approach has proven to substantially enhance user engagement and agency in the creation of generative art. This method merges human creativity with AI by manipulating visual latent spaces, providing users with an intuitive way to influence AI outputs. However, there remain notable discrepancies in the explainability of AI model behavior that require further attention. While incorporating human input helps make AI models more understandable, it does not fully resolve issues of transparency. The Latent Auto-recursive Composition Engine (LACE) offers a practical and intuitive platform that empowers artists with greater control over the AI model, thereby deepening their ownership in the creative process.

Our findings suggest that while time estimation is linked to task complexity and the learning curve associated with the tool, it does not necessarily translate to enhanced ownership. Additionally, a high level of satisfaction with AI control correlates with increased engagement in the creative process. Conversely, when users experience a disconnect between their input and the output, they often relinquish much of the creative effort to the AI. Although visual quality is crucial and significantly impacts the user experience, it is not the sole defining characteristic of art. The intrinsic value of art encompasses a spectrum of attributes like contextual meaning, culture, and the creative process, which current AI models still struggle to fully capture.

In the realm of generative AI, the value of art appreciation is significantly enriched by human-AI collaboration rather than by solely relying on AI with prompt engineering. This collaborative approach not only fosters a more meaningful interaction with the creative process but also enhances the authenticity and appreciation of the resulting artwork.

Section 5.2 Future Work

Building on the insights from this study, it is essential for future research to broaden the participant pool beyond Dartmouth College students to include individuals with a wider range of traditional art backgrounds and technological proficiency. Such expansion is crucial to enhance the generalizability of the findings across different artistic and cultural contexts.

Further refinement is required in the application of the Human-AI Parallel Model (see 2.1.4). The current use of the LACE interface primarily for text and parameter inputs suggests that it may not fully facilitate user engagement. Future iterations of the model should aim to introduce multi-modality inputs that allow artists greater control, thereby better supporting artistic workflows.

Additionally, the Workflow V: Visual Reference Model (see 2.1.5) warrants in-depth exploration to assess its role in using visual references to augment the artistic process, particularly its impact on artistic autonomy and creativity. Research should investigate how artists at varying levels of digital proficiency, from novices to seasoned professionals, utilize this workflow. Conducting targeted experiments across various artistic disciplines, such as painting, graphic design, and concept art, will help determine if Workflow V, as an auxiliary tool, enhances artists' sense of ownership and satisfaction.

By addressing these points and focusing on refining interactions between artists and technology, the goal is to foster a deeper integration of human creativity and artificial intelligence in the art world, ultimately enhancing both the artistic process and the appreciation of art.

Bibliography

[1] Krea.ai. `https://www.krea.ai/`. Accessed: April 18, 2024.

[2] RunwayML. `https://runwayml.com/`. Accessed: [2024-04-30].

[3] Stability AI. Improved autoencoders: sd-vae-ft-ema. `https://huggingface.co/stabilityai/sd-vae-ft-ema`, n.d. 2023. Accessed: [Your access date here].

[4] asagi4. Experimental cads implementation for comfyui. `https://github.com/asagi4/ComfyUI-CADS`, 2024. Accessed: 2024-04-20.

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022.

[6] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting, 2022.

[7] Lucas Bellaiche, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli. Humans versus ai: whether and why we prefer human-created compared to ai-created artwork. *Cognitive Research: Principles and Implications*, 8(1):42, 2023.

[8] Tracy Cassidy. The mood board process modeled and understood as a qualitative design research tool, 11 2011.

[9] Robert Chamberlain, Chris Mullin, Brett Scheerlinck, and Johan Wagemans. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2):177–192, 2018.

[10] Sougwen Chung. Drawing operations. `https://sougwen.com/project/drawing-operations`, 2015-2018. Accessed: 2024-04-23.

[11] Sougwen Chung. Flora rearing agricultural network (f.r.a.n.). `https://sougwen.com/project/florarearingagriculturalnetwork`, 2020-2021. Accessed: 2024-04-23.

[12] Harold Cohen. The further exploits of aaron, painter. *Stanford Hum. Rev.*, 4(2):141–158, jul 1995.

[13] comfyanonymous. Comfyui: The most powerful and modular stable diffusion gui, api and backend with a graph/nodes interface. `https://github.com/comfyanonymous/ComfyUI`, 2024. Accessed: 2024-04-20.

[14] Patricia Cornelio, Patrick Haggard, Kasper Hornbaek, Orestis Georgiou, Joanna Bergström, Sriram Subramanian, and Marianna Obrist. The sense of agency in emerging technologies for human–computer integration: A review. *Frontiers in Neuroscience*, 16, 2022.

[15] Arthur Danto. The artworld. *The Journal of Philosophy*, 61(19):571–584, 1964.

[16] Theresa Rahel Demmer, Corinna Kühnapfel, Joerg Fingerhut, and Matthew Pelowski. Does an Emotional Connection to Art Really Require a Human Artist? Emotion and Intentionality Responses to AI- Versus Human-Created Art and Impact on Aesthetic Experience. *Computers in Human Behavior*, 148:107875, 2023.

[17] Mohammed Diab et al. Stable diffusion prompt book. OpenArt, 11 2022. Available online at `https://openart.ai/promptbook`.

[18] George Dickie. Defining art. *American Philosophical Quarterly*, 6(3):253–256, 1969.

[19] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[20] Victor Gallego. Personalizing text-to-image generation via aesthetic gradients, 2022.

[21] Gabriel Goh. Decoding the thought vector. `https://gabgoh.github.io/ThoughtVectors/`, 2022. Accessed: [2024-04-23].

[22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[25] iamkaikai. amazing_logos_v4. `https://huggingface.co/datasets/iamkaikai/amazing_logos_v4`, 2023. Accessed: 2024-04-26.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[27] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023.

[28] James W. Moore and Sukhvinder S. Obhi. Intentional binding and the sense of agency: A review. *Consciousness and Cognition*, 21(1):546–561, 2012. Beyond the Comparator Model.

[29] NimaNzrii. comfyui-photoshop: Photoshop node inside of comfyui, send and get data from photoshop. `https://github.com/NimaNzrii/comfyui-photoshop`, 2023. Accessed: 2024-04-21.

[30] Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models, 2023.

[31] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018.

[32] Python Software Foundation. Tkinter — python interface to tcl/tk. `https://docs.python.org/3/library/tkinter.html`, 2024. Accessed: 2024-04-21.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[34] Kazimierz Rajnerowicz. Human vs ai test: Can we tell the difference anymore? `https://www.tidio.com/blog/human-vs-ai-test`, March 2024. Accessed: April 2024.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[36] Robin Rombach, Patrick Esser, et al. Stable Diffusion v1-5: A latent text-to-image diffusion model, June 2022.

[37] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling, 2024.

[38] VALADI K JAGANATHAN. efficiency-nodes-comfyui: A collection of comfyui custom nodes. `https://github.com/jags111/efficiency-nodes-comfyui`, 2024. Accessed: 2024-04-20.

[39] Karl Wallick. Generative processes: Thick drawing. *International Journal of Art & Design Education*, 31(1):19–29, 2012.

[40] CALEB WARD. Photoshop vs midjourney inpainting — midjourney inpainting tutorial. `https://www.youtube.com/watch?v=gnzcM0xraPs`, 2023.

[41] Daniel Wegner, Betsy Sparrow, and Lea Winerman. Vicarious agency: Experiencing control over the movements of others. *Journal of personality and social psychology*, 86:838–48, 06 2004.

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[43] Morris Weitz. The role of theory in aesthetics. *The Journal of Aesthetics and Art Criticism*, 15(1):27–35, 1956.

[44] Tom White. Sampling generative networks, 2016.

[45] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models, 2022.

[46] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths, 2024.

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.