Enhancing AI Image-Generation: The Impact of Human-in-the-Loop on Image Preference Alignment and Creativity

LINXI XIE, NYU Shanghai, China
ZHUORAN CHEN, NYU Shanghai, China

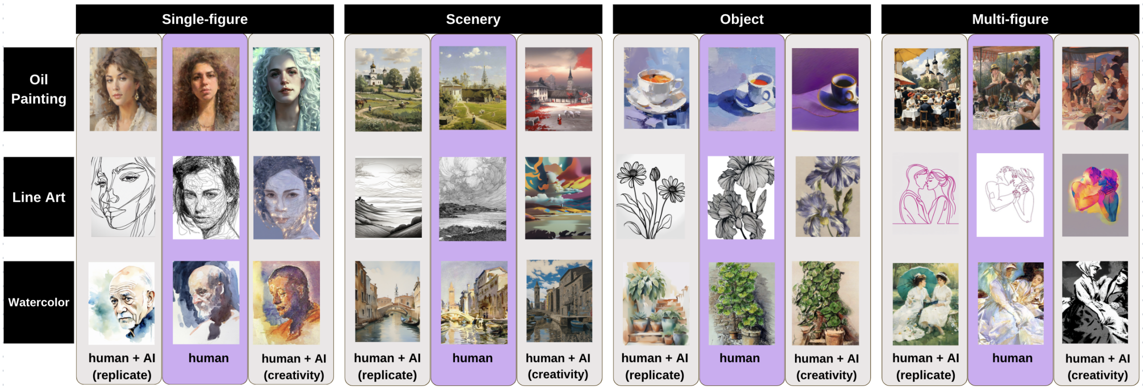


Fig. 1. We present 12 reference images created by real artists (middle), each showcasing different styles and content. Human–AI collaboration is employed both to replicate these images (left) and to enhance their creativity (right).

AI image-generation tools allow users to create novel images tailored to their preferences. This paper examines how human collaboration with these tools helps align outputs with user preferences and support more creative results. We design two tasks to study both the process and outcome of human–AI interaction for image generation. The first task focuses on how humans describe images through content and style controls, and the second evaluates whether human participation can enhance the aesthetic value of original images. We analyze the effectiveness of human prompt modifications across image categories and show how these adjustments improve resemblance to reference images. We further study how human involvement affects the perceived aesthetic quality of generated images. Our findings pinpoint where human participation provides the most benefit and suggest directions for improving generative models for human-in-the-loop image generation.

Additional Key Words and Phrases: image generation, human-in-the-loop, prompt engineering, creativity, controllability, aesthetics, LoRA

1 INTRODUCTION

As Artificial Intelligence (AI) continues to evolve, its role has expanded far beyond automating routine tasks, becoming a powerful tool for enhancing human creativity across industries like music, filmmaking, and literature [1, 17, 24]. Among the different forms of AI-generated content (AIGC), image-generation is particularly noteworthy for its vibrant outputs, as well as its accessibility. Over recent years, platforms like Civitai [1] and LiblibAI [2], alongside other AI image-generation communities, have experienced rapid growth, driven by increased user engagement with generative techniques.

[1] https://www.civitai.com.
[2] https://www.liblib.art.

Authors' addresses: Linxi Xie, lx2154@nyu.edu, NYU Shanghai, Shanghai, China; Zhuoran Chen, zc2745@nyu.edu, NYU Shanghai, Shanghai, China.

As the popularity of AI image-generation grows, research has increasingly focused on enhancing human interaction with these tools to provide users with greater control over the creative process. One of the most widely adopted methods is the use of text-based prompts, which allow users to shape the AI-generated images by describing the desired outcome in words [6]. This practice, commonly referred to as "prompt engineering," has emerged as a key factor in optimizing the performance of text-to-image models [19]. Well-crafted prompts have a significant impact on the final output, helping to align the AI's creations more closely with the user's vision [15]. Beyond text prompts, more advanced models such as LoRA [28] and tools like IP-Adapter [26] and ControlNet [27] offer even greater flexibility by allowing users to control style and structure, thus broadening the creative possibilities in AI-generated image production.

Along with the improvement in user control of image-generation process, evaluating the aesthetics of AI-generated output has also become a major research focus. Studies consistently show that AI-generated art is rated lower in quality compared to human-made art, revealing a persistent bias [3, 12]. This has led to criticism that the creative potential of AI is frequently over-hyped [21], with AI-created and human-created works not being perceived as equal in artistic value [11]. This disparity raises a crucial question: How can AI-generated art achieve the same level of impact as human-made artwork? To address the aesthetic limitations of AI-generated images, various deep neural network architectures and models have been utilized to improve AI-generated art [14]. Research primarily focused on optimizing algorithms and developing advanced network architectures to address this issue [5]. However, relying solely on technical advancements makes it difficult to pinpoint which specific aesthetic aspects need improvement. Few studies have tackled this challenge from a human-centered perspective, focusing on how the creative process can be improved by identifying the weaknesses in the generative pipeline. This approach is essential because human involvement plays a pivotal role in enhancing the aesthetic quality of AI-generated content, offering insights that machines alone cannot provide [10].

To bridge this gap, our study explores the impact of human-AI collaboration on image control and the aesthetic value of the output. We also examine how human intervention contributes to improving image quality and identify the aspects that benefit most from human participation. We focus on two key aspects of AI image-generation: user controllability and aesthetic value. To explore these dimensions, we designed two tasks: The first task investigates how users manipulate content (via prompt refinement) and style (via LoRA models) to create images that align with their intention. This task is designed to simulate real-world scenarios where users have a conceptual target in mind and seek to achieve it through AI tools. The second task examines whether users can enhance the aesthetic quality of the real image by using a style transfer workflow, specifically through IPAdapter, where the original image serves as a guide.

Our study aims to answer the following research questions:

(1) RQ 1: To what extent does human-in-the-loop involvement enhance the controllability of AI-generated images in aligning with a reference image, and which aspects of alignment benefit most from human intervention?
(2) RQ 2: How does human-in-the-loop involvement enhance the aesthetic quality of the original images, and in what aspects does human intervention make improvements?

Through these tasks, our findings reveal the critical aspects where human participation enhances image quality and offer insights into how existing generative models, such as Stable Diffusion XL, can be improved for more effective human-in-the-loop interactions.

2 RELATED WORK

2.1 AI as a Catalyst for Human Creativity

By collaborating with humans in creative processes, AI fosters novel ideas and artistic content across various fields [17]. For entertainment, the 2016 film Sunspring was entirely written by AI, and the recent album Hello World marked the first AI-produced music album [1]. In writing, Stojanovic demonstrated that AI enhances creative writing by providing new inspirations, improving consistency, and refining grammar [24]. In visual arts, text-to-image generators like Midjourney, Stable Diffusion, and DALL-E have revolutionized digital artwork creation by automating artistic processes [29]. Research showed that these text-to-image generators increase human creative productivity by 25% and enhance artwork value, with a 50% higher likelihood of receiving a favorite per view [29].

These studies suggest that people tend to hold negative attitudes toward AI-generated art. To address this, our research explores whether integrating human input into the AI generation process can reduce this bias by allowing users to collaborate with AI in shaping the style and outcome of generated images.

2.2 Bias in AI-generated Art

As AI tools become more popular for creating art, assessing the aesthetics of AI-generated artwork has emerged as a key area of research. Studies consistently show that AI-generated art is rated lower than human-made art in terms of quality [3, 4, 12, 16], highlighting a bias favoring human involvement in creativity. Chiarella et al. [5] confirmed this bias, revealing that skepticism toward AI art stems from perceptions of effort and narrative depth. However, positive attitudes toward AI can sometimes mitigate this bias.

These studies suggest that people tend to hold negative attitudes toward AI-generated art. To address this, our research explores whether integrating human input into the AI generation process can reduce this bias by allowing users to collaborate with AI in shaping the style and outcome of generated images.

2.3 Advancements in User Controls for AI-generated Art

With the rise of image generative models, methods like prompt engineering have emerged to help users create images aligned with their preferences [6, 18, 22]. Oppenlaender [19] introduced six prompt modifiers to guide models, while Goloujeh et al. [15] explored how users refine prompts based on intent. Clarisó et al. [6] proposed a domain-specific language (DSL) for platform-independent prompts, ensuring adaptable results across AI systems. Beyond text prompts, tools like IP-Adapter [26] use target images for guidance, ControlNet [27] refines specific features, and LoRA [28] fine-tunes models with small datasets. The "MultiDiffusion" framework [2] generates diverse, high-quality images based on user-defined controls.

Building on these methods, our research integrates prompts, IP-Adapter, ControlNet, and LoRA, offering users multiple ways to control the image generation process. Additionally, we introduce AI-generated prompts using a state-of-the-art image captioning model to help users refine their prompts, combining human insight with machine efficiency to improve the quality and alignment of generated images.

3 METHODS

3.1 Study Design

3.1.1 Image-Generation Task. We designed two tasks to study human-AI collaborative image-generation. The first task aims to assess whether humans can effectively control the content, structure, and style of output images (see

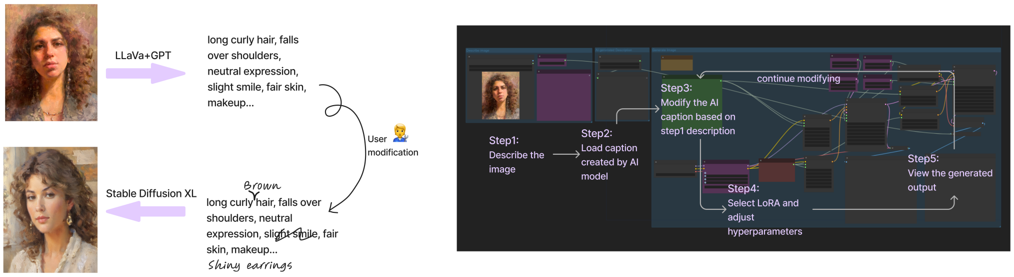

LLaVa+GPT
long curly hair, falls over shoulders, neutral expression, slight smile, fair skin, makeup...
User modification
Brown
long curly hair, falls over shoulders, neutral expression, ~~slight smile~~, fair skin, makeup...
Shiny earrings
Stable Diffusion XL
Step1: Describe the image
Step2: Load caption created by AI model
Step3: Modify the AI caption based on step1 description
continue modifying
Step4: Select LoRA and adjust hyperparameters
Step5: View the generated output

Fig. 2. Overview of Task 1 (left): The reference image is first captioned by LLaVa-NeXT [13], and the user then modifies the prompt to create a more accurate description. The modified prompt is then fed into Stable Diffusion XL [20] to generate the output. Default workflow for Task 1 (right): Describe the reference image, load AI-generated captions, and iteratively refine the prompt based on the output image.

Figure 2). To achieve this, we asked users to replicate an artist's image using a provided workflow. Users were first asked to describe the reference image in their own words. An AI-generated caption, produced by a state-of-the-art image captioning model, was then provided for comparison. Participants were tasked with modifying this caption to more accurately describe the reference image. Importantly, users were only allowed to use the prompt as a guide, not the reference image itself, as the goal was to study the effectiveness of prompt modification. However, they were free to choose LoRA models and adjust hyperparameters as needed. Additionally, users could further refine the prompts based on the output images to achieve better alignment with the target image.

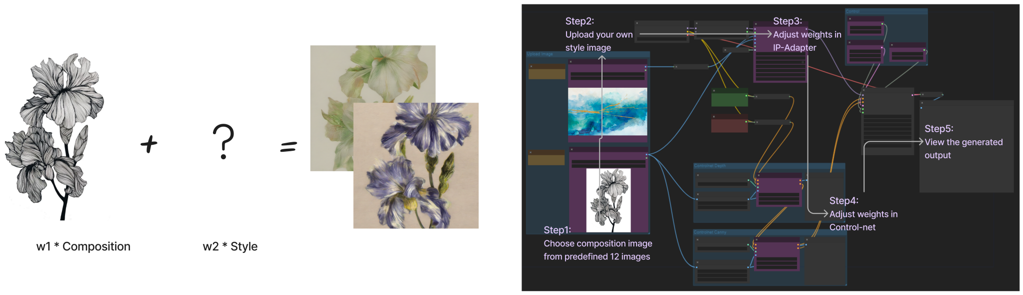
The second task is to explore how effectively humans can enhance the aesthetic value of images (see Figure 3). To allow for creative freedom, minimal restrictions are placed on the generation process. The only requirement is that the major content of the image must be preserved, as this is essential for studying aesthetic improvements. Therefore, users were required to use the reference image as a content guide. We provided users with a basic IP-Adapter workflow that enables the use of a single-image LoRA for style transfer on the reference image. Additionally, users were free to adjust various hyperparameters, such as the weight of the style and content guides, and may employ any additional workflows to refine the output according to their preference.

We used ComfyUI [3] for building the image-generation workflow due to its flexibility in adjusting hyperparameters and its efficient integration of multiple workflows, including LoRAs.

3.1.2 *Evaluation of Generated Image.*

We used Streamlit [4] to build our rating interface (see Figures 4 and 5). For Task 1, which focuses on testing the effectiveness of prompt modification, we established a control group. In this setup, images were generated using AI-generated captions, user-selected hyperparameters, and user-selected LoRA models. The only difference between the control group and the user-refined group was the modification of the prompt. Each real image was paired with two generated images—one based on the human-refined prompt and one based on the unmodified AI-generated caption. This resulted in 12 groups of images. In the evaluation interface, raters viewed three images at once: the real image, labeled as "Reference Image", and the two generated images labeled as "Image 1" and "Image 2". Raters need to vote for the image that best matches the reference, choosing from Style, Content, and Structure as reasons for their choice.

[3] https://github.com/comfyanonymous/ComfyUI
[4] https://github.com/streamlit/streamlit



Step1: Choose composition image from predefined 12 images

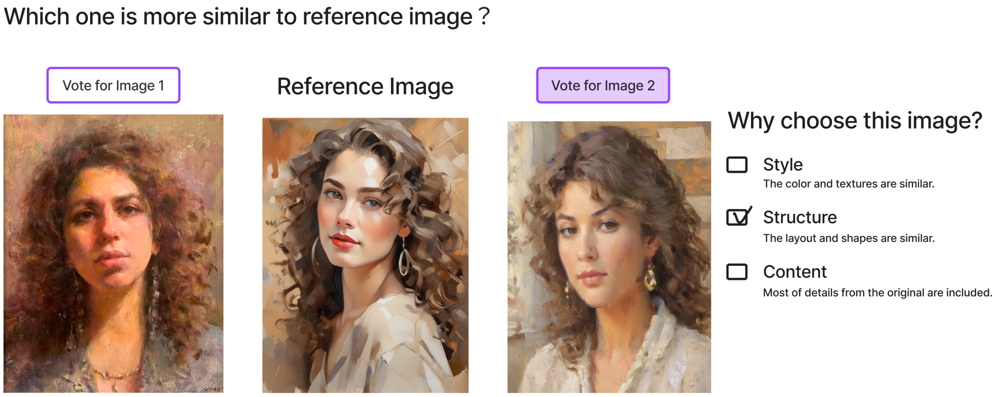
Step2: Upload your own style image

Step3: Adjust weights in IP-Adapter

Step4: Adjust weights in Control-net

Step5: View the generated output

Fig. 3. Overview of Task 2 (left): Choose a composition image from the predefined set of 12 and select any style image based on user preference. Adjust the weights for the composition (w1) and style (w2) guides to control the final generated output. Default workflow for Task 2 (right): Select a composition image, upload a style image, and adjust hyperparameters for IP-Adapter and ControlNet.



Which one is more similar to reference image ?

Why choose this image?

☐ Style
The color and textures are similar.

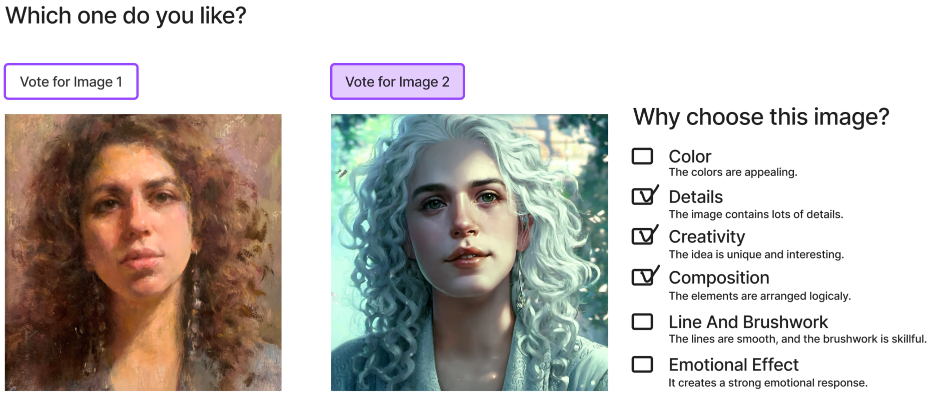
☑ Structure
The layout and shapes are similar.

☐ Content
Most of details from the original are included.

Fig. 4. Task 1 Voting Interface: Users compare Image 1 and Image 2 to the reference image, pick the most similar one, and choose the reasoning behind their decision from the following options: Style, Structure, or Content.

For Task 2, the evaluation followed a comparable structure, with images grouped into 12 sets. Each real image was paired with two generated images, and users were tasked with comparing two images from each set at a time, with all combinations sampled equally. To ensure fair and consistent comparisons, we used the Elo rating system, which dynamically adjusted image ratings based on user preferences. This method ensured that the ratings fairly reflected the relative quality of the images. Users were also asked to provide reasons for their choices, selecting from six criteria: Color, Details, Creativity, Composition, Line and Brushwork, and Emotional Effect.

To prevent user fatigue and maintain high-quality evaluations, each rater was assigned only one task to rate, and users were directed to the task with fewer ratings to ensure a balanced distribution between the two tasks. Additionally, to avoid bias caused by differences in image size or cropping positions between generated and reference images, all images were resized to the same dimensions by centering and cropping outward based on the smallest width and height ratio within the group. To further maintain rating consistency, we prioritized sampling pairs with fewer ratings to ensure balanced coverage and approximately the same number of ratings per image.



Which one do you like?

Why choose this image?

☐ Color
The colors are appealing.

☑ Details
The image contains lots of details.

☑ Creativity
The idea is unique and interesting.

☑ Composition
The elements are arranged logicaly.

☐ Line And Brushwork
The lines are smooth, and the brushwork is skillful.

☐ Emotional Effect
It creates a strong emotional response.

Fig. 5. Task 2 Voting Interface: Users vote for the image they prefer, then select the reason for their choice from the following criteria: Color, Details, Creativity, Composition, Line and Brushwork, or Emotional Effect.

3.2 Participant Recruitment

For the image-generation task, we recruited 11 participants through social media, specifically targeting individuals with backgrounds in design or experience with AI-generated content. This group consisted of 9 females and 2 males, aged between 18 and 30, with expertise ranging from Computer Science to Interaction Media Art. Among them, 4 participants are frequent designers, and 10 participants are familiar with ComfyUI.

In the image evaluation stage, we expanded recruitment by sharing a public link to the rating interface. A total of 57 participants rated the images generated in Task 1, with 24 having experience in AI-generated content (AIGC). For Task 2, 52 participants provided ratings, 26 of whom had AIGC experience.

3.3 Image Selection

We selected 12 reference images from Google Arts & Culture [5] and Pinterest [6], ensuring a diverse representation of artistic styles and content. Specifically, the images consist of three distinct styles: oil painting, watercolor, and line art, and four content categories: single-figure, scenery, object, and multi-figure. To maintain a focus on aesthetic evaluation and minimize bias, we deliberately excluded images depicting religious themes, historical events, or content that could cause discomfort.

3.4 Metric and Model Selection

We used DreamSim, a perceptual metric introduced at NeurIPS 2023, to evaluate image similarity based on human visual perception. Unlike traditional pixel-based metrics, DreamSim focuses on mid-level attributes such as layout, object positioning, and semantic content, capturing subtleties in image perception more effectively. Trained on synthetic data, DreamSim has shown strong generalization to real images and outperforms previous metrics and large vision models in tasks like image retrieval and reconstruction [9].

[5] https://artsandculture.google.com/
[6] https://www.pinterest.com/

Model Type	Blip	Deepseek	Moondream2	UForm-Gen	LLaVa-NeXT 13b	LLaVa-NeXT 13b + GPT-4o
Average similarity across seeds	0.5406	0.4450	0.4551	0.4556	0.4448	0.4514
Average standard deviations across seeds	0.1251	0.1056	0.0915	0.1027	0.0722	0.0733

Table 1. Comparison of different image captioning models based on average similarity and standard deviation scores. The result is computed by DreamSim metric. The results demonstrate that LLaVa-NeXT 13b achieved the lowest DreamSim score and standard deviation, indicating strong and robust performance in generating captions.

This metric is applied in two ways: (1) evaluating the captioning capabilities of text-to-image models by comparing the generated images against reference images, and (2) providing an early-stage evaluation in Task 1 to predict which images human viewers are likely to perceive as most similar to the reference. DreamSim offers us an initial quantitative assessment of image similarity, which we subsequently confirm through human rating scores.

To identify the state-of-the-art model for image captioning, we evaluated five models: BLIP [23], Deepseek-vl-7b-chat [8], Moondream2 [25], UForm-Gen [7], and LLaVa-NeXT 13b [13]. For each model, we generated captions for 12 reference images. Using these captions, we employed Stable Diffusion XL [20] to create images based on 10 consistent random seeds. To assess the performance, we measured the similarity between the reference and generated images using DreamSim. We then calculated the average similarity score and standard deviation across the 10 seed sets for each model. Since a lower DreamSim score indicates better performance, the results showed that the LLaVa-NeXT 13b model not only achieved the lowest DreamSim score but also had the lowest standard deviation, indicating strong and robust performance across different seeds (see Table 1). Based on these findings, we selected LLaVa-NeXT 13b as the optimal model for image captioning.

Additionally, since Stable Diffusion XL can only process 77 tokens at a time, we observed that the captions generated by LLaVa-NeXT 13b consistently exceeded this limit, averaging 153 tokens per image. To address this, we used GPT-4o mini to shorten the captions and format them as a sequence of short phrases for improved readability. After refinement, the average caption length was reduced to 46.5 tokens per image.

3.5 Data Analysis

To evaluate the effectiveness of human-in-the-loop methods, we compared images across two main categories: those generated with human-refined prompts versus those generated without, and user-generated images versus real images. Our analysis focused on two key outcomes: (1) the number of wins in each category, and (2) the reasons contributing to each image category's evaluation. We further categorized images based on content and style.

In Task 1, the number of wins was determined by counting the images with the higher votes. In addition to user ratings, we employed DreamSim to calculate similarity scores for both the user-refined and control groups against the reference images, providing an objective measure of image quality. We also analyzed the percentage distribution of the reasons for each vote. By comparing the percentage changes in reasons before and after refinement, we identified specific areas where user involvement improved image outcomes.

In Task 2, we applied the same process to compare user-generated images with real images. We ensured proportional sampling of image groups (one real and two user-generated) and retained Elo scores for each image. A win was counted for the real image if it had no higher-rated user-generated competitors, and vice versa. We then compared the percentage changes in voting reasons between the real and user-generated images. To maintain fairness, we aggregated the results of the two user-generated images into a single representation (1:1 ratio). This allowed us to identify which aspects of image quality were enhanced by user involvement.

To further explore the effectiveness of human-AI collaboration, we grouped the 12 images by both content and style, analyzing whether improvements varied across content types (e.g., single-figure, multi-figure, scenery, object) and art styles (e.g., oil painting, watercolor, line art). We further compared data changes within these groups to identify patterns that revealed areas where human input improved or failed to improve the quality of the generated images compared to those from human artists.

4 RESULTS

4.1 Prompt Modification is Effective

We compared images generated with unmodified prompts to those generated with modified prompts, assessing the impact of prompt refinements on image alignment. The evaluation using Dreamsim as metric revealed that the average similarity score improved from 0.435 (before modification) to 0.404 (after modification), indicating an improved alignment with the reference images. And the number of images with lower DreamSim scores rose from 3 to 9, further proved this enhancement.

Furthermore, the results from the user ratings showed that the number of winning images increased from 3 to 9 following prompt modifications. In terms of category distribution, the percentage of winning images for Style, Content, and Structure remained relatively stable, with slight shifts observed: Style (36.99% to 36.48%), Content (36.94% to 36.00%), and Structure (26.06% to 27.51%).

4.2 Prompt Modification Enhances Structure Alignment for Single Figure Image

Image Type	Reason	Before (%)	After (%)
Pic#1 (single figure + oil painting)	Structure	0.00	20.37
Pic#9 (single figure + watercolor)	Structure	24.14	38.24
Pic#5 (single figure + line art)	Structure	28.57	33.33

Table 2. Percentage of vote for Structure on single-figure images, both before and after prompt modifications. The percentages reflect how much 'Structure' contributed to the overall reasons for choosing each image, indicating an increasing performance on structure alignment after changes to the prompts.

After conducting an image-wise analysis, we found that in the content-based groups, single-figure images consistently showed significant improvements in structure alignment after prompt modifications, with gains as high as 20.37% (see Table 2). This indicates a notable enhancement in structural quality for simpler compositions. However, for more complex content types (multi-figure, scenery, and object), the improvements were less consistent, suggesting that prompt modifications have a more pronounced effect on simpler, single-figure compositions. In the style-based groups, no clear pattern emerged, indicating that prompt modifications did not significantly enhance the art style of the generated images.

The structural alignment of Pic #1 showed the largest improvement, increasing by 20.37% after prompt modifications. A closer comparison of the prompts before and after modification revealed that some structure-related changes were

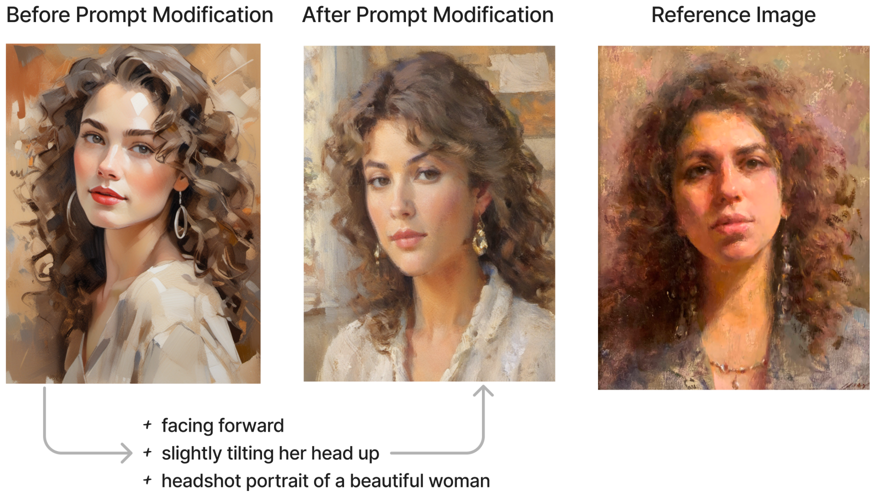

Before Prompt Modification
After Prompt Modification
Reference Image
+ facing forward
+ slightly tilting her head up
+ headshot portrait of a beautiful woman

Fig. 6. By adding structure-related phrases into the initial prompt, the user-refined Picture 1 (middle) achieves better structural alignment with the reference image (right).

effective, while others were not (see Figure 6). For example, including the phrase "headshot portrait" made the figure appear larger, while "facing forward" caused the body to align more directly toward the front. However, the face remained in a side-facing position. Additionally, phrases like "slightly tilting her head up" did not result in any noticeable changes in the output.

The results from Task 1 indicated that generative models like Stable Diffusion XL need to improve their responsiveness to nuanced prompt inputs, particularly in handling subtle adjustments. While broader instructions like "facing forward" successfully influenced body alignment, finer details such as head orientation ("slightly tilting her head up") were not reflected, highlighting a gap in interpretative precision. Enhancing the model's sensitivity to minor structural and stylistic tweaks would enable more effective human-in-the-loop interactions, giving users greater control over outputs and ensuring that even small prompt adjustments lead to meaningful changes in the final image.

4.3 Human Involvement Enhances Color and Creativity, but Falls Short in Line and Brushwork

Image Type	Color	Details	Creativity	Composition	Line and Brushwork	Emotional Effect
Real Image	19.55%	16.99%	11.11%	10.69%	25.63%	16.02%
User-Generated	27.42% ↑	16.11%	14.67%	10.92%	14.51% ↓	16.37%

Table 3. Percentage distribution of reasons for selecting winning images from Task 2 across various criteria. User-generated images show a higher preference for color and creativity compared to real images, but real images are favored for line and brushwork. The results suggest that while AI tools excel in enhancing creativity and color, they struggle to replicate the detail and craftsmanship of human art. (An arrow indicates a significant change.)

When asked, "Which image is better?" only 4 out of 12 user-generated images were preferred over real paintings, revealing that the aesthetics of user-generated images fail to surpass real images.

A deeper analysis of the reasons behind image preferences shows that color and creativity are more frequently appreciated in user-generated images (see Table 3), with selection rates of 27.42% for color and 14.67% for creativity, compared to 19.55% and 11.11% for real paintings, respectively. However, real paintings dominate in terms of line and brushwork, achieving a selection rate of 25.63%, the highest in all categories.

This suggests that AI tools, while capable of enhancing creativity—especially in areas like color—struggle to capture the detailed precision of line and brushwork as seen in traditional paintings. Even with human input, this supports earlier research that AI-generated art does not surpass real paintings in overall preference [3, 12].

4.4 Human Involvement Enhances Color and Creativity in Scenery and Object-Based Images

Image Name	Reason	Real (%)	User (%)
Pic#2 (scenery + oil painting)	Color	24	34.05
Pic#6 (scenery + line art)	Color	3.45	27.18
Pic#10 (scenery + watercolor)	Color	27.18	34.53
Pic#3 (object + oil painting)	Creativity	9.8	12.5
Pic#7 (object + line art)	Creativity	6.45	9.3
Pic#11 (object + watercolor)	Creativity	3.85	13.32

Table 4. Percentage of votes for Color and Creativity in scenery and object-based images (Real vs. User-Generated). The result suggests that user-generated images often excel in Color and Creativity.

Image Name	Reason	Real (%)	User (%)
Pic#5 (line art + single-figure)	Color	0	28.1
Pic#6 (line art + scenery)	Color	3.45	27.18
Pic#7 (line art + object)	Color	9.68	34.64
Pic#8 (line art + multi-figure)	Color	0	25.22

Table 5. Percentage distribution of votes for Color in line art images (Real vs. User-Generated). This table provides insight into how style adjustments impact the perception of color.

A closer examination of different image categories reveals that the selection rate for Color increases significantly in scenery images, with an increment ranging from 7.35% to 23.73% (see Table 4). This trend suggests that user-generated images of scenery tend to be more vibrant and visually impactful, likely due to stylistic choices that idealize the environment. For images featuring object, Creativity sees a notable rise in user-generated images, with increments ranging from 2.75% to 9.47%. This indicates that users tend to produce more imaginative and inventive content when focusing on object.

For line art images, the most significant changes are observed in color (see Table 5). All of the images show over a 20% increase in the percentage of votes for color when comparing real images to user-generated images. This outcome is expected, as the real images rely on simple lines, while user interaction enhances them by adding more intricate and vibrant color schemes.

The results from Task 2 show that while generative models like Stable Diffusion XL are highly effective at producing colorful and creative images, they need improvement in capturing finer artistic elements such as line quality, texture, and brushwork. To strike a balance, these models should enhance the display of fine details without allowing an overemphasis on color and creativity to compromise realism and artistic sophistication.

5 LIMITATIONS AND FUTURE WORK

Through surveys, we found that users had difficulty controlling object positions through prompts. This highlights a clear direction for future work: improving the responsiveness of image-generation to positional descriptions in prompts. To better understand how users modify prompts to gain control, Task 1 should be expanded to trace users' prompt adjustments. This investigation will focus on identifying what makes prompt modifications successful and what causes failures in generating high-quality images, by evaluating both the results and the process.

Besides, AI-generated art still falls short of human-created art in terms of aesthetics. The reasons for this gap remain unclear, and further investigation is needed to determine if factors such as the "AI Effect" [5] contribute to these differences. Future research should focus on enhancing generative models like Stable Diffusion XL by improving control over key elements such as texture, lighting, and object positioning. Incorporating more sophisticated style modifications and expanding prompt customization could provide valuable insights into the model's potential. Continuous user feedback and comparative studies across different user groups are crucial for refining Stable Diffusion's algorithms to better meet user expectations and improve the aesthetic quality of AI-generated images.

Lastly, future studies should involve a larger dataset, exploring the use of larger captioning models, testing various methods for modifying image attributes, and including a diverse group of participants could help better distinguish effective approaches. Future work should explore how individuals unfamiliar with AI-generated content (AIGC) navigate these systems compared to more experienced users. Identifying strategies to enhance the user interface could improve the accessibility and usability of AIGC tools. By addressing these areas, the experiment's basic setup can be enhanced to achieve more comprehensive results.

6 CONCLUSION

This paper explores the impact of human-AI collaboration on aligning outputs with preferences and enhancing the aesthetics of original images. Our findings demonstrate that while structural alignment remains challenging to achieve with prompt modification, certain image categories, such as single-figure images, are more amenable to structural adjustments. Although human involvement effectively enhances aspects like creativity and color, the generated output still cannot outperform the original images—real paintings created by human artists—in human rating evaluations. This is primarily due to the superior detail and finer brushwork inherent in real paintings, which AI-generated images struggle to replicate. Our results suggest that generative models require improvements in two key areas: (1) responsiveness to nuanced prompt inputs, as evidenced by Task 1 where models like Stable Diffusion XL faced difficulties with subtle prompt adjustments (e.g., head orientation), and (2) the ability to produce fine artistic elements such as line quality and texture, which remains a challenge as demonstrated in Task 2.

REFERENCES

[1] G. Amato, M. Behrmann, F. Bimbot, B. Caramiaux, F. Falchi, A. Garcia, J. Geurts, J. Gibert, G. Gravier, H. Holken, H. Koenitz, S. Lefebvre, A. Liutkus, F. Lotte, A. Perkis, R. Redondo, E. Turrin, T. Vieville, and E. Vincent. 2019. AI in the media and creative industries. *arXiv* (2019). arXiv:1905.04175 [cs.AI] https://arxiv.org/abs/1905.04175

[2] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel. 2023. MultiDiffusion: Fusing diffusion paths for controlled image generation. *arXiv* (2023). arXiv:2306.09344 [cs.CV] https://arxiv.org/abs/2306.09344

[3] L. Bellaiche, R. Shahi, M. H. Turpin, et al. 2023. Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications* 8, 1 (2023), 42. https://doi.org/10.1186/s41235-023-00499-6

[4] R. Chamberlain, C. Mullin, B. Scheerlinck, and J. Wagemans. 2018. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts* 12, 2 (2018), 177–192. https://doi.org/10.1037/aca0000136

[5] S. Chiarella, G. Torromino, D. Gagliardi, D. Rossi, F. Babiloni, and G. Cartocci. 2022. Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior* 137 (2022), 107406. https://doi.org/10.1016/j.chb.2022.107406

[6] R. Clarisó and J. Cabot. 2023. Model-driven prompt engineering. In *Proceedings of the 2023 ACM/IEEE 26th International Conference on Model Driven Engineering Languages and Systems (MODELS)*. IEEE, 47–54. https://doi.org/10.1109/MODELS58315.2023.00020

[7] Unum Cloud. 2024. uform-gen2-qwen-500m. https://huggingface.co/unum-cloud/uform-gen2-qwen-500m. Accessed: 2024-09-13.

[8] DeepSeek-AI. 2024. DeepSeek VL-7B Chat. https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat. Accessed: 2024-09-13.

[9] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint* (2023). arXiv:2306.09344 [cs.CV]

[10] Aaron Hertzmann. 2018. Can Computers Create Art? *Arts* 7, 2 (2018). https://doi.org/10.3390/arts7020018

[11] Joo-Wha Hong and Nathaniel Ming Curran. 2019. Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence. 15, 2s, Article 58 (jul 2019), 16 pages. https://doi.org/10.1145/3326337

[12] M. Khan and J. Näsström. 2023. *How AI images and human-created digital art are evaluated and affected by attribution knowledge.* Ph. D. Dissertation. KTH Royal Institute of Technology. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-333666 Dissertation.

[13] LLaVA-HF. 2024. LLaVA v1.6 Vicuna 13B HF. https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf. Accessed: 2024-09-13.

[14] Anne-Sofie Maerten and Derya Soydaner. 2024. From paintbrush to pixel: A review of deep neural networks in AI-generated art. arXiv:2302.10913 [cs.LG] https://arxiv.org/abs/2302.10913

[15] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 183, 13 pages. https://doi.org/10.1145/3613904.3642861

[16] M. Mazzone and A. Elgammal. 2019. Art, creativity, and the potential of artificial intelligence. *Arts* 8, 1 (2019), 26. https://doi.org/10.3390/arts8010026

[17] M. Mrak. 2019. AI gets creative. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV '19)*. Association for Computing Machinery, 1–2. https://doi.org/10.1145/3347449.3357490

[18] J. Oppenlaender. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek '22)*. Association for Computing Machinery, 192–202. https://doi.org/10.1145/3569219.3569352

[19] J. Oppenlaender. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* (2023), 1–14. https://doi.org/10.1080/0144929X.2023.2286532

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=di52zR8xgf

[21] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3334480.3382892

[22] Laria Reynolds and K. McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–7.

[23] Salesforce. 2024. BLIP Image Captioning Large. https://huggingface.co/Salesforce/blip-image-captioning-large. Accessed: 2024-09-13.

[24] L. Stojanovic, V. Radojcic, S. Savic, D. Sarcevic, and A. S. Cvetković. 2023. The influence of artificial intelligence on creative writing: Exploring the synergy between AI and creative authorship. *Journal of Creative Writing Studies* 12, 1 (2023), 70–74.

[25] VikhyatK. 2024. MoonDream2. https://huggingface.co/vikhyatk/moondream2. Accessed: 2024-09-13.

[26] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. 2023. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint* (2023). arXiv:2308.06721 [cs.CV]

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV] https://arxiv.org/abs/2302.05543

[28] M. Zhong, Y. Shen, S. Wang, Y. Lu, Y. Jiao, S. Ouyang, D. Yu, J. Han, and W. Chen. 2024. Multi-LoRA composition for image generation. *arXiv preprint* (2024). arXiv:2402.16843 [cs.CV]

[29] E. Zhou and D. Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (2024), page052. https://doi.org/10.1093/pnasnexus/pgae052

A APPENDIX: TASK 1 VOTING RESULTS

Image Name	Vote	Style	Structure	Content
1_after	41	36	13	16
1_before	5	4	1	1
2_after	44	28	24	32
2_before	2	1	1	2
3_after	40	30	14	26
3_before	6	2	2	3
4_after	35	18	16	22
4_before	11	2	5	10
5_after	39	26	24	18
5_before	7	2	3	3
6_after	18	15	4	9
6_before	27	9	18	18
7_after	6	4	2	4
7_before	39	28	13	17
8_after	36	16	21	28
8_before	9	7	4	3
9_after	24	9	14	15
9_before	21	13	7	12
10_after	32	23	14	16
10_before	14	8	6	5
11_after	18	5	10	7
11_before	28	18	12	16
12_after	39	23	20	19
12_before	7	2	3	3

Table 6. User voting information for all 12 image pairs in Task 1. Each pair includes an image generated with a human-refined prompt ("after") and one without human modification ("before").

B APPENDIX: TASK 2 VOTING RESULTS

Image Name	Elo	Color	Details	Creativity	Composition	Line and Brushwork	Emotional Effect
1st Image Pair							
1_real	1173.687741	14	10	3	7	16	13
1_user1	862.5588998	4	5	3	3	4	2
1_user2	963.7533588	7	0	5	3	4	4
2nd Image Pair							
2_real	1154.383252	21	19	3	12	14	19
2_user1	900.3447225	8	6	4	4	2	5
2_user2	945.2720252	9	3	2	3	1	3
3rd Image Pair							
3_real	1126.220513	21	13	6	5	13	6
3_user1	916.408626	10	2	1	4	4	5
3_user2	957.3708607	8	3	5	4	3	2
4nd Image Pair							
4_real	1126.405762	15	18	3	6	12	9
4_user1	969.2771471	17	4	5	6	5	9
4_user2	912.4776366	1	5	2	2	2	1
5nd Image Pair							
5_real	1064.283712	0	7	12	8	16	11
5_user1	1047.447413	8	8	2	5	10	7
5_user2	888.2688752	7	3	4	2	1	2
6nd Image Pair							
6_real	1022.79489	2	12	5	4	12	5
6_user1	1082.522617	20	12	8	8	11	15
6_user2	894.6824929	4	1	4	2	1	3
7nd Image Pair							
7_real	942.8803796	3	7	3	7	11	4
7_user1	1044.097105	14	7	1	3	7	7
7_user2	1013.022515	14	11	9	1	9	5
8nd Image Pair							
8_real	1005.157521	0	2	10	6	9	7
8_user1	979.2840333	12	2	8	3	4	8
8_user2	1015.558445	10	10	14	5	6	11
9th Image Pair							
9_real	1074.663329	13	5	6	8	13	4
9_user1	887.1312248	8	2	5	1	3	4
9_user2	1038.205446	7	9	6	5	10	10
10th Image Pair							
10_real	1021.94108	12	4	6	3	10	11
10_user1	1030.417558	16	4	7	3	6	10
10_user2	959.697343	12	9	3	4	6	5
11th Image Pair							
11_real	922.3696558	8	4	2	2	10	6
11_user1	1089.190186	10	15	2	8	18	2
11_user2	988.4401583	10	8	10	5	5	5
12th Image Pair							
12_real	1152.843797	24	17	2	5	14	13
12_user1	944.6520655	3	6	4	5	5	8
12_user2	902.5041377	5	0	3	2	3	5

Table 7. User voting information for all 12 image groups in Task 2. Each group includes a reference image ("real") and two user-generated images ("user1" and "user2") based on the reference.

C IMAGE CREDITS

- Aaron Coberly. *Untitled*. Available at: https://www.aaroncoberly.com/watercolor-gallery
- Britt Nicole. *Hugging People*. Available at: https://www.pinterest.com/pin/4011087173917286/
- Faberge Julia. *Iris*. Available at: https://www.pinterest.com/pin/1407443627331124/
- Franklin Booth. *The Shoreline*. Available at: https://www.pinterest.com/pin/281543720015597/
- John S. Sargent. *Simplon Pass: Reading*. Available at: https://artsandculture.google.com/asset/simplon-pass-reading/EAEb0-10WMOkCQ
- John Singer Sargent. *Venice San Giuseppe Castle Bridge*. Available at: https://watercoloracademy.com/watercolor-masters/john-singer-sargent
- Kris Trappeniers. *A Lady*. Available at: https://www.pinterest.jp/pin/77405687324278983/
- Mary Qian. *Susan*. Available at: https://artzline.com/mary-quan-susan/
- Paul Cézanne. *Geraniums*. Available at: https://artsandculture.google.com/asset/geraniums-0013/ZQGyv2e689pUuQ
- Pierre-Auguste Renoir. *Luncheon of the Boating Party*. Available at: https://artsandculture.google.com/asset/luncheon-of-the-boating-party-0001/mgHsTKDNJVzPAg
- Rachel Petruccillo. *Cup and Shadow*. Available at: https://www.pinterest.com/pin/95208979614099866/
- Vasily Polenov. *Moscow Patio*. Available at: https://artsandculture.google.com/asset/moscow-patio/EwFgf3K9mAQuyg