Reflective Human-Machine Co-adaptation for Enhanced Text-to-Image Generation Dialogue System

Anonymous ACL submission

Abstract

Today's image generation systems are capable of producing realistic and high-quality images. However, user prompts often contain ambiguities, making it difficult for these systems to interpret users' potential intentions. Consequently, machines need to interact with users multiple rounds to better understand users' intents. The unpredictable costs of using or learning image generation models through multiple feedback interactions hinder their widespread adoption and full performance potential, especially for non-expert users. In this research, we aim to enhance the user-friendliness of our image generation system. To achieve this, we propose a reflective human-machine co-adaptation strategy, named RHM-CAS. Externally, the Agent engages in meaningful language interactions with users to reflect on and refine the generated images. Internally, the Agent tries to optimize the policy based on user preferences, ensuring that the final outcomes closely align with user preferences. Various experiments on different tasks demonstrate the effectiveness of the proposed method.

1 Introduction

Generative artificial intelligence has demonstrated immense potential in facilitating economic development by helping optimize creative and non-creative tasks. Models such as DALL·E 2 (Ramesh et al., 2021), IMAGEN (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and Muse (Chang et al., 2023) have achieved this through their capability to produce unique, convincing, and lifelike images and artwork from textual descriptions (Gozalo-Brizuela and Garrido-Merchan, 2023). Despite the considerable progress achieved, there remains substantial potential for improvement, particularly in generating higher-resolution images that more accurately reflect the semantics of the input text and in designing more user-friendly interfaces (Frolov et al., 2021). Many models find it hard to accurately comprehend the nuanced intentions behind human instructions, often leading to a mismatch between user expectations and model outputs.

Moreover, the impact of certain adjustments to variables on the final image output is not always straightforward, posing a significant challenge for non-expert users who haven't systematically learned prompt engineering courses. The intricacy involved in comprehending and manipulating these variables presents a substantial obstacle for individuals without a technical background. Furthermore, given the same input text, the model may still generate images with substantially different content or layouts, where aspects such as background, color, and perspective can vary. In such instances, the user must engage in multiple trials, and acquiring an image that meets their specific requirements can depend significantly on chance.

To address these challenges, we introduce an innovative dialogic approach designed to enhance the user experience for non-professional users. Within this dialogic interaction process, we posit the existence of a latent generative objective in the user's mind. A single image may represent the user's latent and unconscious generative goal. By iteratively querying the user, we can progressively elicit more detailed descriptions, with the ultimate aim of producing an image that closely aligns with the user's underlying intent. Figure 1 illustrates the operational flow of this project as interacted by the users. This approach is inspired by the concept of human-in-loop co-adaptation (Reddy et al., 2022), where the model evolves alongside user feedback to better align with user expectations. Our main contributions are:

- We delve into human-machine interaction methods within image generation tasks, guiding users to effectively create images that reflect their intentions and preferences.

- We introduce an enhanced Text-to-Image dialogue-based Agent, which leverages both external interactions with users and internal reflections to enhance its performance.
- Application across the general image and fashion image generation demonstrates the versatility and potential value of our approach.

2 Related work

Text-Driven Image Editing Framework

Recent advancements in text-to-image generation have focused on aligning models with human preferences, using feedback to refine image generation. Studies range from Hertz et al. (Hertz et al., 2022)'s framework, which leverages diffusion models' cross-attention layers for high-quality, prompt-driven image modifications, to innovative methods like ImageReward (Xu et al., 2024), which develops a reward model based on human preferences. These approaches collect rich human feedback (Wu et al., 2023; Liang et al., 2023), from detailed actionable insights to preference-driven data, training models for better image-text alignment and adaptability (Lee et al., 2023) to diverse preferences, marking significant progress in personalized image creation.

Ambiguity Resolution in Text-to-Image Generation

From visual annotations (Endo, 2023) and model evaluation benchmarks (Lee et al., 2024) to autoregressive models (Yu et al., 2022) for rich visuals, along with frameworks for abstract (Liao et al., 2023) and inclusive imagery (Zhang et al., 2023), the text-to-image field is advancing through strategies like masked transformers (Chang et al., 2023), layout guidance (Qu et al., 2023) without human input, and feedback mechanisms (Liang et al., 2023) for quality. The TIED framework and TAB dataset (Mehrabi et al., 2023) notably enhance prompt clarity through user interaction, improving image alignment with user intentions, thereby boosting precision and creativity.

Human Preference-Driven Optimization for Text-to-Image Generation Models

Zhong et al. (Zhong et al., 2024) significantly advance the adaptability of LLMs to human preferences with their innovative contributions. Zhong et al.'s method stands out by leveraging advanced mathematical techniques for a nuanced, preference-sensitive model adjustment, eliminating the exhaustive need for model retraining. Xu et al. (Xu et al., 2024) take a unique approach by harnessing vast amounts of expert insights to sculpt their ImageReward system, setting a new benchmark in the creation of images that resonate more deeply with human desires. Together, these advancements mark a pivotal shift towards more intuitive, user-centric LLMs technologies, heralding a future where AI seamlessly aligns with the complex mosaic of individual human expectations.

3 Proposed method

We developed a modular architecture tailored for image generation tasks within multi-turn dialogues. This architecture is designed to facilitate deep introspection of the generation system and effectively guide user interactions. The system comprises several key components: The *Memory* stores the dialogue, denoted as h. The *Summarizer*, denoted as M_S, integrates users' historical dialogue content, and generates a *Prompt*, denoted as P, for image generation. The *Generation Model*, denoted as M_G, is responsible for transforming P into specific images. The *Reflection Block*, denoted as B_R, plays a crucial role. It not only handles the reasoning process (completing tasks in collaboration with the user) but also engages in internal reflection on the model. Within this module, the *Evaluator*, marked as M_E, is tasked with providing a comprehensive description of the generated images. The *Ambiguity Inference* M_{inf} analyses the potential ambiguity and outputs an internal label r. Finally, the *Action*, designated as M_A, displays the image and poses questions to the user. We provide a detailed exposition of this interactive framework, distinguishing between its internal and external workflows.

3.1 External Reflection via Verbal Reflection

The external reflection is contingent on user interactions. When the user presents a new prompt, the agent generates a corresponding image and subsequently reflects on which intents to inquire about based on that image. This interactive process is termed Human-Machine Reflection (HM-Reflection).

Memory and Summarizer The historical dialogues between the user and the agent are stored in the *Memory*, while the *Summarizer* M_S generates

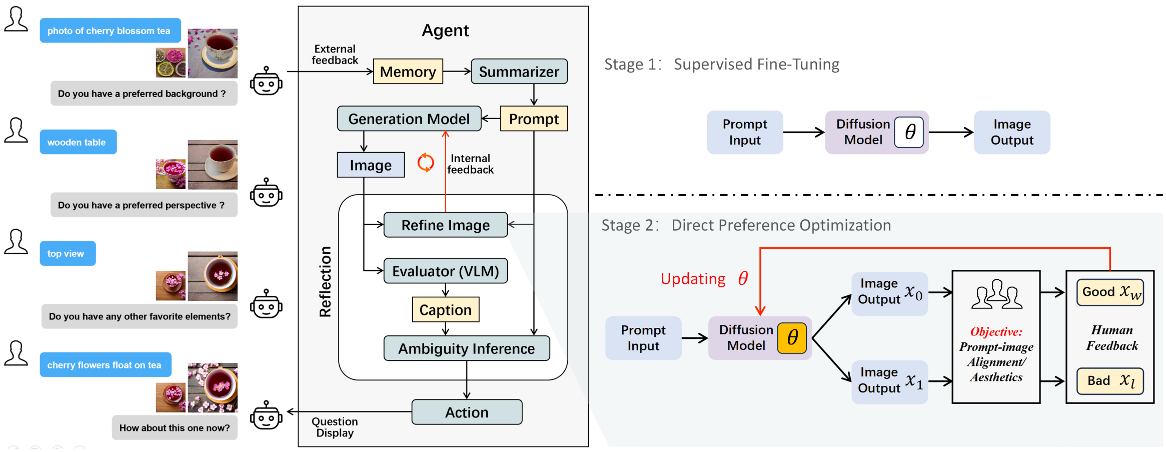


Figure 1: Proposed framework of Enhanced Text-to-Image Reflexion Agent. The Generation Model can learn user preferences by Direct Preference Optimization.

the prompt for controlling image generation based on these historical dialogues. Let h represent the historical dialogues, t represent the current time, w_t represent the current user's response, and P_t represent the internal prompt used for image generation. The entire process can be expressed with the following formula:

$$P_t = M_S(w_t, h). \tag{1}$$

Generation Model The *Generation Model* M_G is central to the image generation, creating images based on provided prompts. Besides generating images that align with user intentions, it also incorporates additional details not explicitly mentioned by the user. For the general image generation task, we use the Stable Diffusion model v1.4 (Rombach et al., 2022). Specifically, for the fashion image generation task, we employ a Stable Diffusion XL v1.0 (Podell et al., 2023), fine-tuned on fashion-related datasets. This is because fashion images are generally uniform in layout and demand a richer representation of fine-grained features. Let I_t represent the currently generated image. This process can be expressed as:

$$I_t = M_G(P_t). \tag{2}$$

Evaluator In this interactive reflection framework, the *Evaluator* M_E plays a critical role in assessing the quality of the generated images. The *Evaluator* uses a visual language model (VLM) to describe the image content and generates captions that include aspects such as content, style, and background. We utilize Qwen-VL (7B) (Bai et al., 2023) in the general image generation task and ChatGPT 4.0 (OpenAI, 2023) in the fashion image creation task, as the VLM evaluator. The generated captions are represented as C_t, where C_t encompasses N aspects of the description.

$$C_t = M_E(I_t),\ C_t = \left\{C_t^1, C_t^2, \ldots, C_t^N\right\}. \tag{3}$$

Inference and Action By comparing the similarity between multiple captions C_t and the prompt P_t, the *Ambiguity Inference* Model M_{inf} identifies which contents are expected by the user and which are randomly generated, and output an Ambiguitiy label r_t. Based on the detected ambiguities r_t, the *Action* M_A asks the user for more detailed information. Question q_{t+1} can be selected from a predefined list of questions or generated by a large language model (LLM) based on the captions and prompts.

$$r_t = M_{inf}(C_t, P_t), \tag{4}$$

$$q_{t+1} = M_A(C_t, r_t). \tag{5}$$

The entire process of external reflection has been formalized into Algorithm 1.

3.2 Internal Reflection via Direct Preference Optimization

An efficient intelligent interaction system not only provides effective feedback and guidance to users but also has the ability to self-reflect. As illustrated in Figure 1, the Agent features a *'Refine Image'* step that optimizes the model or output results. After generating multiple images, users can mark the ones they prefer. The Agent then learns user preferences from this feedback to produce images that better align with user preferences. We employ a reinforcement learning method D3PO (Yang et al., 2023) for preference learning, which directly learns

Algorithm 1 External reflection via Verbal Reflection

1: Initialize Agent: M_S, M_G, M_E, B_R, M_A
2: **while** dialog **do**
3: User input words: w_t
4: Store w_t into Memory h
5: Summarizer M_S generates Prompt P_t
6: Generation Model M_G generates Image I_t
7: Reflection B_R:
8: Evaluator M_E generates Caption C_t
9: Inference Ambiguity r_t
10: Action M_A generates Question q_{t+1}
11: Store q_{t+1} into Memory h
12: **end while**

---

from user feedback without the need for training a reward model. This functionality is designated as Tool 1. Additionally, we offer Tool 2, which checks the quality of generated images and regenerates those that do not align with the corresponding prompt.

Tool 1: Direct Preference Optimization (DPO) Figure 1 illustrates the method of internal reflection via DPO. In Stage 1, the generation model undergoes supervised fine-tuning to adapt to a specific generation task. In Stage 2, a certain amount of preference feedback is accumulated through multiple interactions with the user. This feedback is then used to optimize the model, resulting in more personalized outputs. The optimization method employed is D3PO (Yang et al., 2023), which expands the theoretical DPO into a multi-step MDP (Markov Decision Process) and applies it to diffusion models.

Given two image samples, the user selects the image they prefer, denoted as x_w, while the other sample can be represented as x_l. Using the same weight, initialize a reference model π_{ref}, and a target model π_θ. During the denoising process, the diffusion model takes a latent s as input and outputs a latent a. Based on the probability of π_{ref}, the overall loss of the D3PO algorithm gives:

$$\mathcal{L}(\theta) = -\mathbb{E}\left[\log \rho\left(\beta \log \frac{\pi_\theta(a^w \mid s^w)}{\pi_{\text{ref}}(a^w \mid s^w)} - \beta \log \frac{\pi_\theta(a^l \mid s^l)}{\pi_{\text{ref}}(a^l \mid s^l)}\right)\right] \tag{6}$$

Here, β is the temperature parameter that controls the deviation of $\pi_\theta(a|s)$ and $\pi_{ref}(a|s)$. θ is the parameter of the target model.

Algorithm 2 Tool 1: Direct Preference Optimization with D3PO

Require: preferred samples and the other: x_w, x_l and Corresponding Latent: s_w, s_l, a_w, a_l; number of training epochs N; number of prompts per epoch K
1: Copy a pre-trained diffusion model $\pi_{ref} = \pi_\theta$. Set π_{ref} with `requires_grad` to `False`.
2: **for** $n = 1$ to N **do**
3: Training:
4: **for** $k = 1$ to K **do**
5: Update θ with gradient descent using Equation 6
6: **end for**
7: **end for**

Tool 2: Attend-and-Excite The publicly available Stable Diffusion model exhibits issues with *catastrophic neglect*, where the model fails to generate the subjects or attributes from the input prompt. To address this issue in diffusion models and improve text-image alignment, we utilize the A&E algorithm (Chefer et al., 2023).

First, we calculate the CLIP similarity score Sim between the image and prompt. Then, we identify the neglected words by backpropagating the loss function $l = 1 - Sim$. During the process of regenerating the image, we use the A&E method to activate these neglected words. Repeat the above process a certain number of times. This Tool is detailed in Algorithm 3.

Algorithm 3 Tool 2: Attend-and-Excite

Require: Image I_t, Prompt P_t.
1: Initialize $token\_list \leftarrow empty$, Iteration Number N, Threshold k
2: **for** $n = 1$ to N **do**
3: Computing the Similarity of I_t and P_t: $Sim \leftarrow \text{CLIP}(I_t, P_t)$
4: **if** Image is OK: $Sim > k$ **then**
5: **break**
6: **end if**
7: Computing the Objective: $l \leftarrow 1 - Sim$
8: Computing P_t gradient by l: ΔP_t
9: Locate peak value of ΔP_t to get $token\_id$
10: Append $token\_id$ to $token\_list$
11: Regenerate I_t by **A&E**$(P_t, token\_list)$
12: **end for**
13: **return** Image I_t

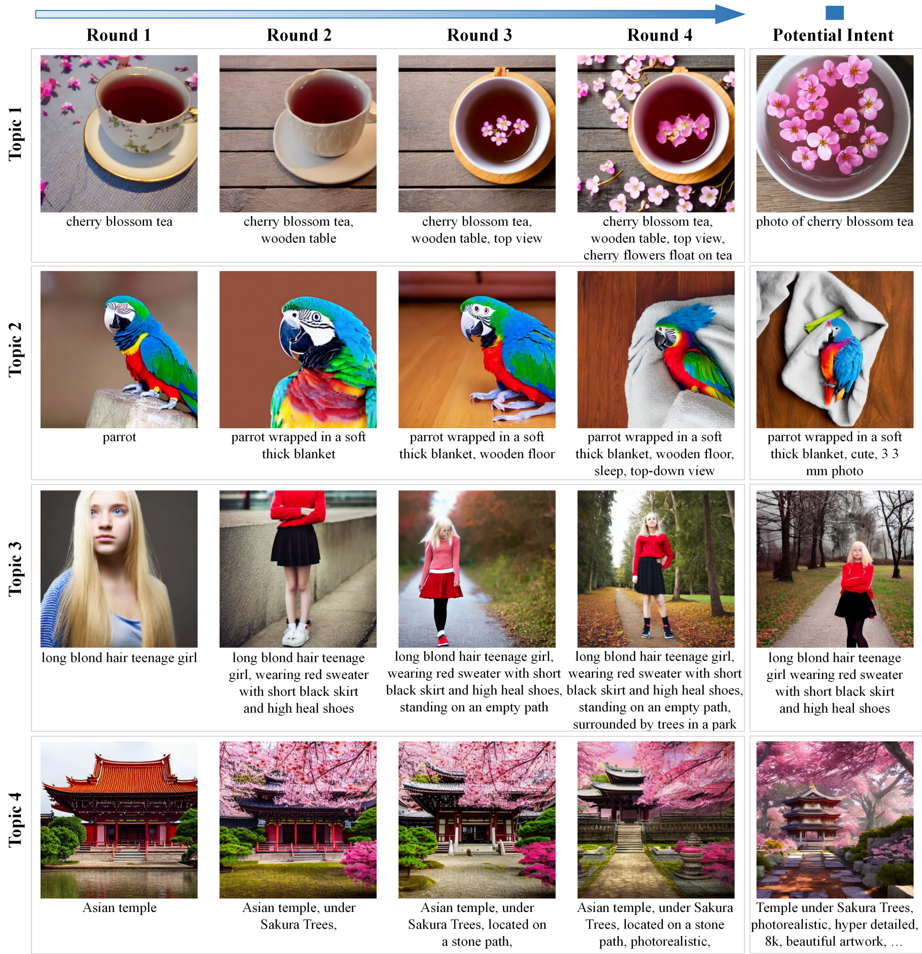


Figure 2: A comparative display of four rounds of image generation based on specific prompts, including cherry blossom tea, a parrot, a teenage girl, and an Asian temple across different rounds.

Table 1: Evaluations of prompt-intent alignment, image-intent alignment and human voting across various methodologies and integrations. Augmentation refers to using LLMs to infer ambiguity and enhance the initial prompt. HM-Reflection is the external reflection of our RHM-CAS. T2I stands for Text-to-Image, and I2I stands for Image-to-Image.

Methods	**Prompt-Intent Alignment**		**Image-Intent Alignment**		**Human Voting**
	T2I CLIPscore	**T2I BLIPscore**	**I2I CLIPscore**	**I2I BLIPscore**	
GPT-3.5 augmentation	0.157	0.145	0.624	0.633	4%
GPT-4 augmentation	0.163	0.152	0.648	0.637	3.2%
LLaMA-2 augmentation	0.112	0.132	0.593	0.571	6%
Yi-34B augmentation	0.101	0.123	0.584	0.560	4.4%
HM-Reflection	0.282	0.281	0.752	0.760	25.5%
HM-Reflection + ImageReward RL	0.292	0.283	0.782	0.776	26.2%
RHM-CAS (Ours)	**0.328**	**0.334**	**0.802**	**0.813**	**30.6%**

Table 2: Multi-dialog (HM-Reflection) ablation experiment with image-to-image similarity scores across different rounds, including SD-1.4, SD-1.5, DALL-E. I2I stands for Image-to-Image.

Multi-dialog	**SD-1.4**		**SD-1.5**		**DALL-E**	
	I2I CLIPscore	**I2I BLIPscore**	**I2I CLIPscore**	**I2I BLIPscore**	**I2I CLIPscore**	**I2I BLIPscore**
Round 1	0.726	0.702	0.722	0.698	0.650	0.673
Round 2	0.757 ($\uparrow$ 0.031)	0.737 ($\uparrow$ 0.035)	0.745 ($\uparrow$ 0.023)	0.724 ($\uparrow$ 0.026)	0.673 ($\uparrow$ 0.023)	0.689 ($\uparrow$ 0.016)
Round 3	0.775 ($\uparrow$ 0.049)	0.762 ($\uparrow$ 0.060)	0.772 ($\uparrow$ 0.050)	0.783 ($\uparrow$ 0.085)	0.690 ($\uparrow$ 0.040)	0.717 ($\uparrow$ 0.044)
Round 4	0.802 ($\uparrow$ 0.076)	0.823 ($\uparrow$ 0.121)	0.788 ($\uparrow$ 0.066)	0.810 ($\uparrow$ 0.112)	0.741 ($\uparrow$ 0.091)	0.735 ($\uparrow$ 0.062)

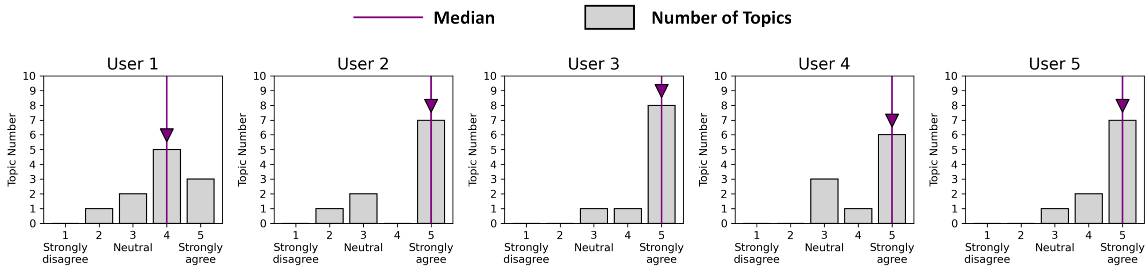


Figure 3: Human Voting for Statement: Multi-turn dialogues can approximate the user's potential intents.

4 Experiment

We explore the application of our proposed Enhanced Text-to-Image Reflexion Agent in two distinct scenarios: general image generation and specific fashion product creation. Due to the different requirements of these applications, adjustments have been made to our approach accordingly. In the experiments, the focus varies between the two tasks. For the general image generation task, we emphasize the effectiveness of our external reflection via verbal reflection. The emphasis of the fashion product creation task is placed on capturing fine-grained features within the images and addressing user preferences.

4.1 Task 1 General Image Generation

The General Image Generation Task, powered by the Enhanced Text-to-Image Reflexion Agent, is designed to enhance the user experience in image creation. Our agent not only generates images based on textual instructions but also engages in dynamic dialogues with users, ensuring the images align more closely with their underlying intentions. This interactivity ensures that the images are not only visually appealing but also meet the content expectations and needs of the users. Moreover, through real-time feedback loops and continuous interaction, the agent guides users and enhances their creative expression, allowing even those with minimal experience to easily produce professional-level images.

4.1.1 Setting

In this task, the process begins with the *Summarizer* generating prompts by aggregating the user's input words. These prompts are then used to generate images. The generated images are subsequently captioned by Qwen-VL (Bai et al., 2023), a Vision-Language Model, covering seven aspects: 'Content', 'Style', 'Background', 'Size', 'Color', 'Perspective', and 'Other'. By comparing the CLIP text similarity scores between the user's historical inputs and each caption, we identify which aspects of the image contain ambiguity. From the three aspects with the lowest scores, one is randomly selected for questioning. The question is displayed, and the user can choose whether to respond.

To quantify the effectiveness of human-in-the-loop image generation, we assumed a reference image as the user's generation target in the experiments. After each image generation, the user responds based on the content of the target image until a certain number of iterations are completed. The similarity between each generated image and the target image is then evaluated to assess the effectiveness of our approach.

4.1.2 Data Collection

We collected those high-scoring image-text pairs from the ImageReward (Xu et al., 2024) dataset, which were gathered from real users. These high-scoring images exhibit excellent visual quality and a high degree of consistency with the original prompts. We excluded samples that were abstract or difficult to understand, as well as those with excessively long input prompts. Ultimately, we obtained 496 samples covering a variety of subjects, including people, animals, scenes, and artworks. And obtained over 2000 prompts from users for image generation. Some of these images also contained content not explicitly mentioned in the original prompts. These reference images served as potential targets for multi-turn dialogue generation, with each sample undergoing at least four rounds of dialogue.

4.1.3 Baseline setup

To demonstrate the effectiveness of our Reflective Human-Machine Co-adaptation Strategy in uncovering users' underlying intentions, we established several baselines. One approach to resolving ambiguity in user prompts is to use Large Language Models (LLMs) to rewrite the prompts. We employed several LLMs to augment the initial prompts, allowing these models to infer the users' intentions. These LLMs include: **ChatGPT-3.5**, **ChatGPT-4** (Achiam et al., 2023), **LLaMA-2** (Touvron et al., 2023), and **Yi-34B** (AI et al., 2024). The relevant experiments are shown in Table 1. Table 1 presents the alignment between the generated prompt and target image, as well as the alignment between the output image and target image. A subjective visual evaluation (Human Voting) was used to select the image result that most closely resembles the target image. All experiments were conducted on four Nvidia A6000 GPUs. The diffusion model SD-1.4 employed the DDIM sampler.

Additionally, we validated the effectiveness of our Multi-dialog (HM-Reflection) approach in uncovering users' underlying intentions by using different generative models. The relevant experiments are shown in Table 2, including **Stable Diffusion (v1.4)**, **Stable Diffusion (v1.5)** (Rombach et al., 2022), and **DALL-E** (Ramesh et al., 2021).

4.1.4 Result Analysis

In Figure 2, we illustrate our reflective human-machine co-adaptation strategy. The rightmost side of the figure shows the target images observed by users during testing, serving as the users' intended generation targets. The four columns of images on the left correspond to the image results and prompt outputs at different dialogue turn. From the visual results, it is evident that by incorporating comprehensive descriptions across the seven aspects, the generated images increasingly align with the target images.

Tables 1 and Table 2 describe the experiments conducted on our collected dataset. Table 1 uses the SD-1.4 as the generative model and Qwen-VL as the evaluator. It first compares the effectiveness of non-human-machine methods (LLM augmentation) in inferring user intent and then evaluates the performance of our multi-dialog approach (HM-Reflection). We compare our RHM-CAS method with a reinforcement learning approach using the feedback of ImageReward model (Xu et al., 2024) to improve the generative model. In Table 1, 'Intent' refers to the target images in the experiments. We use CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) to extract embeddings of prompts and images and measure their similarity scores with the Intent embeddings. Table 1 also includes user votes on which method produced outputs closest to the target images. Compared to other methods, our approach achieved optimal performance. Table 2 shows the effectiveness of multi-dialog (HM-Reflection) in resolving ambiguity across different generative models. As the number of dialog rounds increases, the generated images increasingly resemble the target images, with scores in parentheses indicating the improvement relative to the initial scores. Figure 3 collects the approval ratings from five testers. In these sets of dialogues conducted by each of the five users, we explore whether the users agree that the multi-round dialogue format can approximate the underlying generative target. In most cases, HM-Reflection produces results that more closely align with user intent. Besides, the experiments related to **Tool 2: Attend-and-Excite** are provided in the Appendix D.

4.2 Task 2 Fashion Product Creation

Our second task is fashion product creation, a key application of image generation technology. In the future, generating fashion products like dresses and jackets that users can purchase or customize holds great potential. This approach combines personalization and automation, offering highly customized shopping experiences. Users can generate ideal designs through simple text descriptions, reducing trial and error costs. Brands and designers can quickly test market reactions, lower inventory risks. Overall, image generation technology in fashion has a promising future.

4.2.1 Setting

Fashion product creation is more challenging than general image generation due to higher demands on image quality and diversity. Our Agent system also requires enhanced reasoning and multimodal understanding capabilities. During the experiments, we used ChatGPT 4.0 for reasoning tasks beyond image generation, facilitating multimodal dialogues. More information is available in Appendix B.2.

For image generation, we used the SD-XL 1.0 model for its superior capabilities. We referred to the DeepFashion dataset (Liu et al., 2016) for clothing types and attributes, creating labels for collecting SD-XL 1.0 image samples. These images were cleaned and curated for fine-tuning, resulting in more stable and consistent outputs. The LoRA (Hu et al., 2021) method was used for fine-tuning on four Nvidia A6000 GPUs.

To offer a customized user experience, we

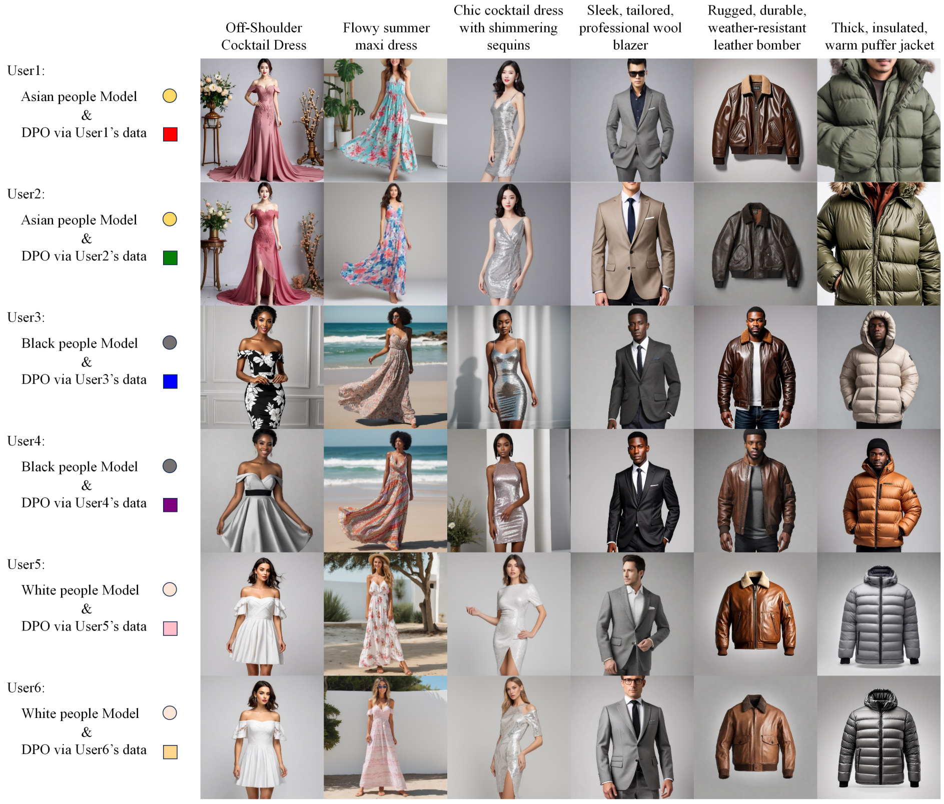


Figure 4: This image showcases a diverse collection of fashion models and outfits, segmented by user preferences or data. Each section highlights different styles of attire, including elegant dresses and professional to casual jackets, modeled by individuals of different ethnic backgrounds.

trained multiple models with different data, allowing users to choose models with different ethnicities. Based on user feedback, the model performs Direct Preference Optimization (DPO). In the DPO process, model parameters are updated after every 40 user feedback instances, repeated three times. The model uses the DDIM sampler for image generation.

4.2.2 Result Analysis

In Figure 4, we display the outputs of six models used by different users, each optimized based on their initial model selections and interaction history. All models generated fashion products from the same prompt using identical seeds, resulting in subtle variations among the products.

We input the same prompt into each of the six models under consistent conditions to produce six sets of fashion items. These products were then processed through Fashion-CLIP (Chia et al., 2022), a version of CLIP fine-tuned for the fashion domain, to obtain their embedding representations, which were visualized in a low-dimensional space using the t-SNE method in Appendix C. The visualization Appendix C Figure 10 shows distinct preference distributions for each user.

Additionally, we had the six testers compare the outputs from models optimized with DPO and those without optimization. As shown in Appendix C Figure 11, in the majority of cases, testers believed that the DPO method improved the model's output results, more aligned with their tastes.

5 Conclusion

In this study, we explored the application of advanced image generation techniques integrated with human-machine interaction frameworks to enhance personalization and visual appeal in both general image generation and fashion product creation. Our Enhanced Text-to-Image Reflection System demonstrated significant capabilities in guiding users to articulate their generative intentions effectively. By leveraging both external interactions and internal reflections, our agent was able to learn from human feedback and align its outputs more closely with user preferences. Future work will focus on integrating finer user feedback mechanisms and broadening the applicability and effectiveness of these technologies in various domains.

6 Limitations

This study, although advanced with the RHM-CAS, has certain limitations. In the interaction process, due to prompts containing multiple high-level descriptions, the image generation model might not fully transform all of them into images. Moreover, the VL model's ability to capture fine-grained details is limited, which may result in inaccurate captions. These cross-modal information transfer processes also lead to errors in information propagation, obstructing the expression of user intent, and thereby affecting communication efficiency. Apart from this, the method is computationally intensive, requiring substantial resources, which may limit its accessibility for users with less powerful hardware. Furthermore, the iterative refinement process, while effective, can be time-consuming. This could potentially lead to user frustration in time-sensitive situations.

Future efforts should aim to enhance computational efficiency and broaden the system's ability to generalize across more diverse inputs, improving usability in real-world applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. 2022. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1).

Yuki Endo. 2023. Masked-attention diffusion guidance for spatially controlling text-to-image generation. *The Visual Computer*, pages 1–13.

Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209.

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2024. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2023. Rich human feedback for text-to-image generation. *arXiv preprint arXiv:2312.10240*.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2023. Text-to-image generation for abstract concepts. *arXiv preprint arXiv:2309.14623*.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14367–14388.

OpenAI. 2023. Chatgpt 4.0. https://www.openai.com/. Accessed: 2024-05-22.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Siddharth Reddy, Sergey Levine, and Anca Dragan. 2022. First contact: Unsupervised human-machine co-adaptation via mutual information maximization. *Advances in Neural Information Processing Systems*, 35:31542–31556.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. 2023. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. 2023. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *arXiv preprint arXiv:2402.02030*.

A Q&A Software Annotation Interface

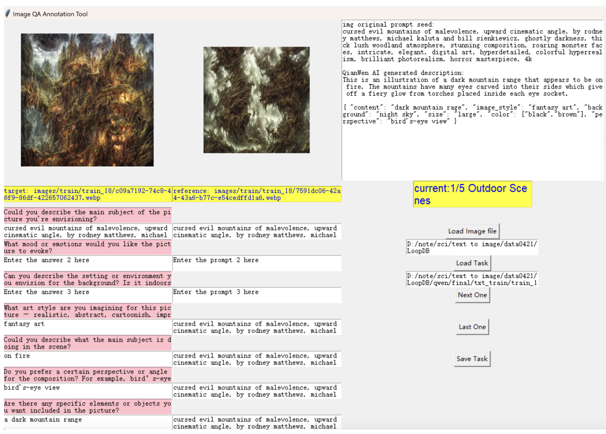


Figure 5: Screenshot of the Q&A software annotation interface.

Image Panel: Two images are displayed side-by-side for comparison or annotation. These images seem to depict artistic or natural scenes, suggesting the software can handle complex visual content.

HTML Code Snippet: Below the images, there's an HTML code snippet visible. This could be used to embed or manage the images within web pages or for similar digital contexts.

Interactive Command Area: On the right, there is an area with various controls and settings:

Current task and image details: Displayed at the top, indicating the task at hand might be related to outdoor scenes. Navigation buttons: For loading new images and navigating through tasks. Annotation tools: Options to add text, tags, or other markers to the images. Save and manage changes: Buttons to save the current work and manage the task details.

A.1 Human annotation instruction

Objective

Accurately describe and tag visual content in images to train our machine learning models.

Steps

1. **Load Image:** Use the 'Load Image' button to begin your task.
2. **Analyze and Describe:**
    - Examine each image for key features.
    - Enter descriptions in the text box below each image.
3. **Tagging:**
    - Apply relevant tags from the provided list to specific elements within the image.
4. **Save Work:** Click 'Save Task' to submit your annotations. Use 'Load Last' to review past work.

Guidelines

- **Accuracy:** Only describe visible elements.
- **Consistency:** Use the same terms consistently for the same objects or features.
- **Clarity:** Keep descriptions clear and to the point.

Support

For help, contact the project manager at [contact information].

Note: Submissions will be checked for quality; maintain high standards to ensure data integrity.

Human annotator information

We invited annotators, users, and testers from university undergraduate and graduate students, including both computer science and non-computer science majors. Compensation was provided based on the amount of work completed.

B RHM-CAS Pipeline Example

B.1 general image generation task pipeline

RHM-CAS uses the Qwen-VL as the evaluator when performing general image generation tasks. Figure 6 presents an example. On the far left is the prompt generated by the Summarizer based on the user's historical dialogues, using the simplest method of phrase stacking for this task. The diffusion model then generates an image based on the current prompt. This image is subsequently described by the Qwen-VL model, which generates captions covering various aspects including "Content," "Image Style," "Background," "Subject Size," "Color," "Perspective," and "Other Aspects." The prompt and the captions are then compared, and a question related to a specific aspect is extracted from the question list. Figure 7 shows a subset of the optional questions from the general image generation task question set.

B.2 fashion product creation task pipeline

When generating fashion products, we attempted to use LLMs to handle all tasks other than image generation. We selected ChatGPT-4 to manage all textual interactions with users and image descriptions, while the generative model used was our fine-tuned Stable Diffusion XL model. As shown in Figure 7, we first initialized several modules based on ChatGPT-4, including *Summarizer*, *Evaluator*, and *Action*. Yellow represents the user's role, while other colors represent different modules of our RHM-CAS. When captioning, the Evaluator provided descriptions from multiple aspects, including 'Appearance,' 'Function,' 'Material,' 'Style,' 'Details,' 'Occasion,' and others. It can be seen that through our RHM-CAS, users can dynamically adjust the generated images and make selections based on recommendations posed by the LLM, allowing even users without prior experience to adapt quickly.

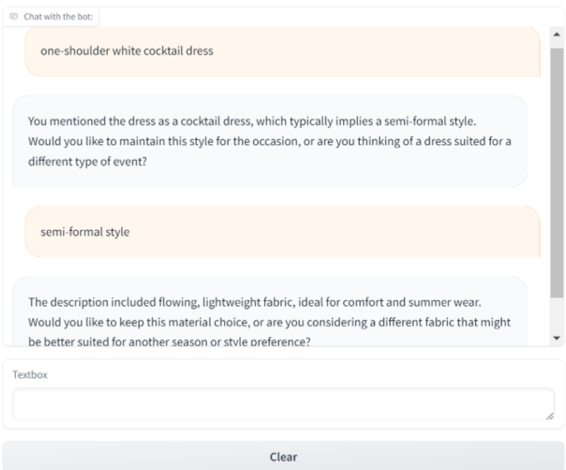
Figure 8 showcases our demo developed based on ChatGPT. The left side of the interface is dedicated to dialogues with users, while the right side generates images in real-time based on the current

Prompt		Image	AI Caption	Reflexion Question	
prompt1:	"cherry blossom tea"		Content: Tea time Image Style: Vintage Background: Tablecloth with flowers Subject Size: Small Color: Pink and white Perspective: Top-down view Other Aspects: Relaxing atmosphere	**question1:**	"Do you have a preferred background ?"
prompt2:	"cherry blossom tea, wooden table"		Content: Tea cup on table Image style: Still life Background: Table top Subject size: Small Color: Red-brown Perspective: Top view Other Aspects: Warm tone	**question2:**	"Do you have a preferred perspective ?"
prompt3:	"cherry blossom tea, wooden table, top view"		Content: Tea with flowers Image style: Overhead view Background: Wooden table Subject size: Small Color: Brown, pink Perspective: Top-down angle Other Aspects: Relaxing atmosphere	**question3:**	"Do you have any other favorite elements?"
prompt4:	"cherry blossom tea, wooden table, top view, cherry flowers float on tea"		Content: Tea time Image Style: Floral tea cup Background: Wooden table Subject Size: Small Color: Pinkish red Perspective: Top view Other Aspects: Cherry blossom flowers	**question4:**	"How about this one now?"

Figure 6: Dialogue Record of General Image Generation, including Prompts, Qwen-VL Captions and Questions.

Question Set	{"Content": "Could you describe what the main subject is doing in the scene ?"} {"Background": "Can you describe the setting or environment you envision for the background? Is it indoors, outdoors, or something abstract ?"} {"Color&Texture": "Could you describe the color or texture of the main subject ?"} {"Style": "What art style are you imagining for this picture, realistic, abstract, cartoonish, impressionistic, etc ?"} {"Size": "Do you have any size requirements for the target?"} {"Perspective": "Do you prefer a certain perspective or angle for the composition? For example, bird's-eye view, worm's-eye view, or eye level?"} {"Others": "Are there any specific elements or objects you want included in the picture ?"}

Figure 7: A subset of the optional questions from the general image generation task question set.



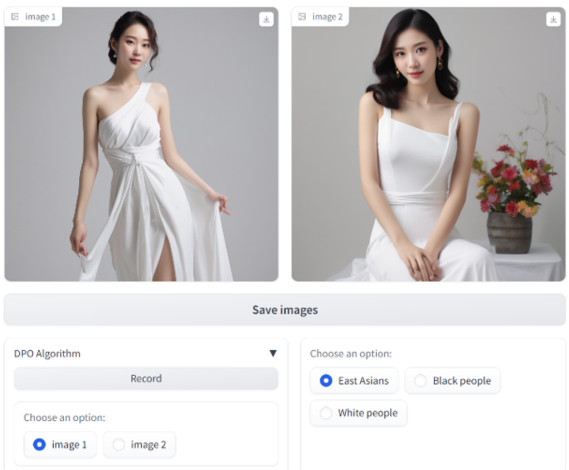


Figure 8: Demo of Fashion Product Creation

User

Initialize Caption Agent: I will give you some fashion images. Please remember to caption the fashion product from 8 aspects (Appearance Function Material Style Details Occasion Brand Fit), using few words in each aspect

Initialize Summary Agent: Remember, each time I give you some words, you'll help me combine them into a complete prompt that I will use to generate an image.

Initialize Action Agent: Remember: Assume a user generates an image through their own prompt, and then a robot generates a caption describing the image. Your role is to ask the user, based on their prompt and the robot's caption of the image, if they want the fashion item as described or prefer a different style. Provide two suggested words as options for their response (do not repeat aspects that have already been inquired about).

Dialog	Agent
white dress Based on the caption describing the white dress as "romantic, off-shoulder gown" under Style, would you prefer the dress in this style, or would you like a different style? Perhaps a **sleeveless** or **high-neck** design?	white dress **Appearance:** Elegant, soft white, flowing silhouette **Function:** Formal wear, primarily bridal **Material:** Lace, tulle **Style:** Romantic, off-shoulder gown **Details:** Lace embroidery, sheer puffed sleeves **Occasion:** Wedding, formal events **Brand:** Not specified **Fit:** Tailored bodice, voluminous skirt SD-XL Output 
sleeveless Considering the caption describes the white dress as having a "V-neckline" under Details, would you like to keep this feature, or would you prefer a different neckline? Maybe a **scoop neck** or **halter top**?	A sleeveless white dress **Appearance:** Sleek, clean lines, white **Function:** Casual or semi-formal wear **Material:** Lightweight fabric, possibly cotton or blend **Style:** Sleeveless dress with a drawstring waist **Details:** Button front, V-neckline **Occasion:** Casual outings, summer events **Brand:** Not specified **Fit:** Relaxed with adjustable waist tie SD-XL Output 
scoop neck Given that the caption mentions the dress is "mid-length" under Style, would you like to keep this length, or would you prefer a different one? Maybe a **knee-length** or **full-length** option?	A sleeveless white dress with a scoop neck. **Appearance:** Minimalist, clean, pure white **Function:** Versatile daywear **Material:** Lightweight, flowy fabric **Style:** Simple, sleeveless, mid-length **Details:** Gathered waistline, no sleeves **Occasion:** Everyday wear, casual outings **Brand:** Not specified **Fit:** Relaxed, comfortable fit SD-XL Output 
full-length Since the caption describes the dress as having a "high-low skirt" under Details, would you like to keep this feature, or are you interested in exploring a different skirt style? Perhaps a **straight hem** or **layered skirt**?	A full-length, sleeveless white dress with a scoop neck. **Appearance:** Sleek, asymmetrical hemline **Function:** Semi-formal or casual elegant wear **Material:** Lightweight, flowing fabric **Style:** Contemporary, sleeveless **Details:** Gathered waist, high-low skirt **Occasion:** Casual outings, semi-formal events **Brand:** Not specified **Fit:** Comfortable bodice with a loose skirt SD-XL Output 

Figure 9: Mode of Fashion Product Creation based on ChatGPT. Special Prompt initialized Each Agent.

conversation. The system presents two images, allowing users to choose the one they prefer, which is then used to optimize the generative model through DPO. Before using the system, users can select different ethnicities in the bottom right corner to initialize the generative model.

C DPO User Study

In the fashion product creation task, we collected feedback from six users and used this feedback to optimize the model through DPO. As shown in Figure 10, under the same random seed conditions, these six models, which have been optimized multiple times, generate images using the same textual input. These images are then fed into the Fashion-CLIP (Chia et al., 2022) model for embedding representation. Finally, these embedding vectors are visualized using the t-SNE method. From the latent space of Fashion-CLIP, it is evident that each of the six models exhibits distinct distribution characteristics.

In addition, we invited these users to evaluate the effectiveness of DPO in Figure 11. Based on their assessments, in most cases, using DPO significantly improved the output performance of the model compared to the unoptimized version.

D Tool 2 ttend-and-Excite Experimrnt

We conducted independent experiments on Algorithm 3 (Tool 2: Attend-and-Excite) using the dataset collected from Task 1. As shown in Table 3, the second row records the usage frequency of Tool 2 as the threshold k varies. When the threshold k is set to 0.72 and 0.7, the usage frequencies are 31.1% and 51.1%, respectively. Correspondingly, the CLIP scores increased by 1.8% and 2.3%, indicating that these settings effectively enhance the alignment between images and text. The iteration number N is set to 3.

E Flawed Example

However, we encountered some suboptimal cases during our experiments. As shown in Figure 12, in the first topic discussing 'Super Mario', the model generated multiple rounds of images based on random noise. As the prompt length increased, the model's understanding of 'Super Mario' gradually diminished, making it difficult to consistently produce a cartoon character. Moreover, the layout of the images was also influenced by the random seed. In some instances, even with added descriptions, it was challenging to obtain images that completely matched the target image, as illustrated in the second topic in Figure 12.

F Potential Risks and Ethical Considerations

The research on image generation based on dialogue systems involves several potential risks that need to be addressed to ensure ethical use and social responsibility.

Firstly, we utilized image generation models from the open-source community. These models have implemented efforts to prevent the generation of misleading or false information. Watermarking techniques have been applied, and strict review mechanisms for content generation have been established to prevent misuse.

Fairness and privacy are also important considerations. The datasets used in this study are based on open-source data, with all user data anonymized and securely stored to protect privacy.

Furthermore, the software programs developed based on these open-source data and models are intended solely for academic research and are not used for commercial purposes.

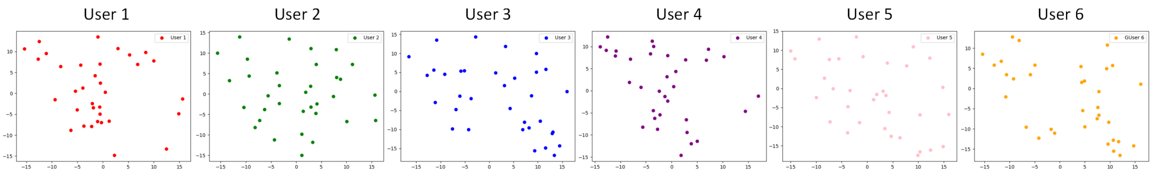


Figure 10: Fashion-CLIP Embeddings of 6 Users visualized with t-SNE

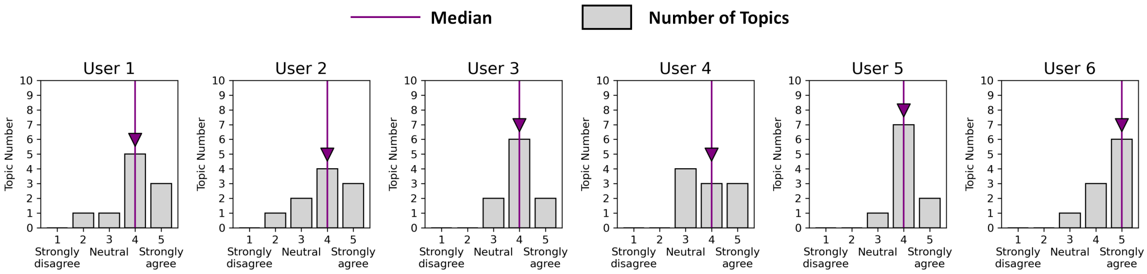


Figure 11: Human Voting for Statement: Direct Preference Optimization can improve generation results.

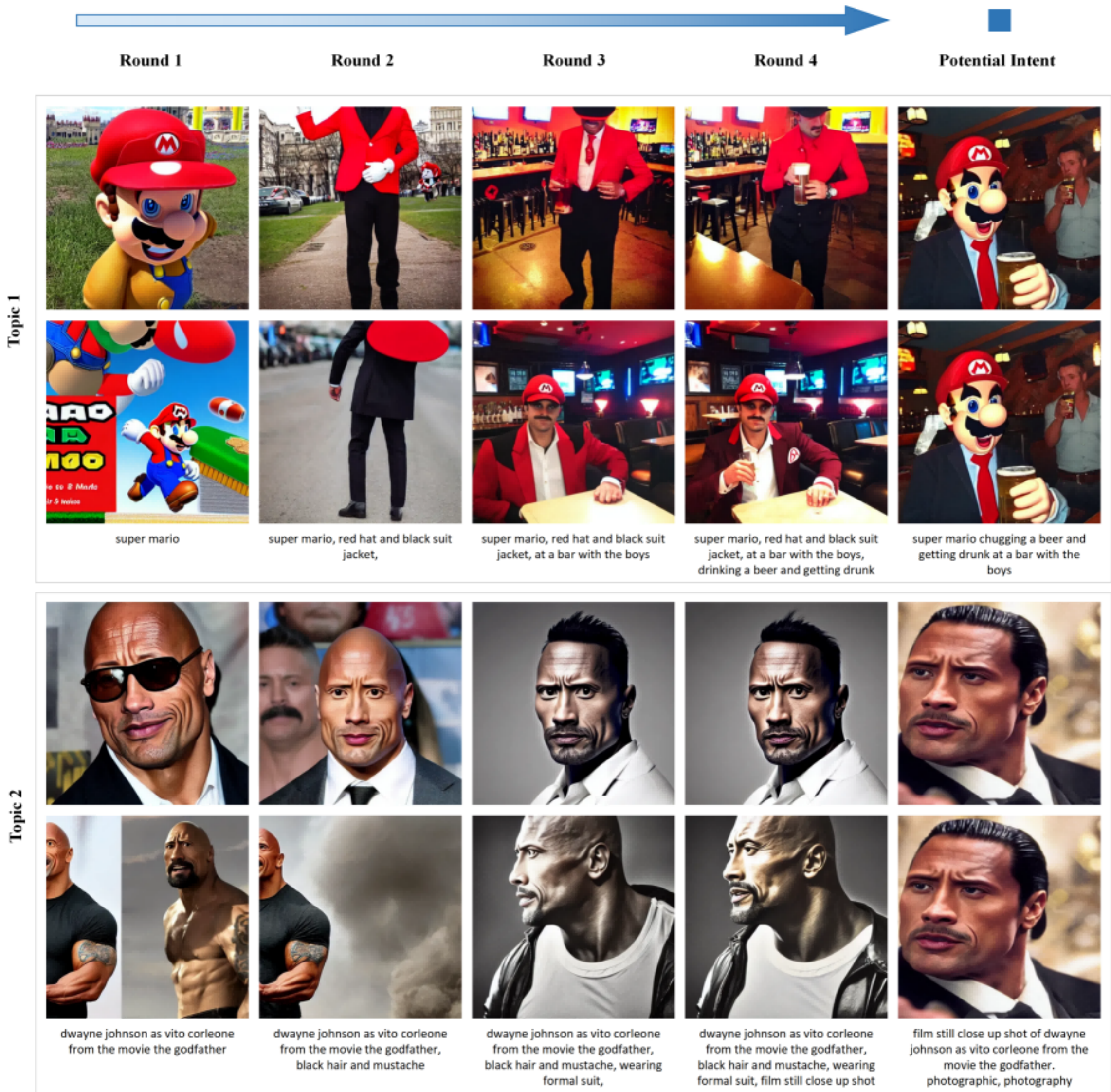


Figure 12: Flawed Case

Tool 2 threshold	0.8	0.75	0.72	0.7	0.68	0.66
Frequency of Usage	0	8.9%	31.1%	51.1%	73%	95.5%
T2I Similarity Improvement	0	0.2%	1.8%	2.3%	2.6%	1.0%

Table 3: Tool 2 usage frequency and T2I Similarity at Different Tool 2 Thresholds