# TDRI: TWO-PHASE DIALOGUE REFINEMENT AND CO-ADAPTATION FOR INTERACTIVE IMAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Today's text-to-image generation technologies have revolutionized the creation of realistic and high-quality images, but they often struggle with the ambiguities in user prompts. To address this, we introduce TDRI: Two-Phase Dialogue Refinement and Co-Adaptation for Interactive Image Generation, a framework designed to enhance iterative image generation through multi-turn dialogues. The system operates in two phases: an initial generation phase that processes user prompts to create base images, and an interactive refinement phase that adapts and optimizes images based on user feedback. Our framework ensures generated outputs continuously adapt to user preferences through iterative dialogue and optimization. Experiments validate that TDRI improves user experience and efficiency, generating high-quality images with fewer iterations, thus streamlining the creative process in various design applications.

## 1 INTRODUCTION

Generative artificial intelligence has shown tremendous potential in driving economic growth by optimizing both creative and non-creative tasks. Advanced models such as DALL·E 3 Betker et al. (2023), Imagen Saharia et al. (2022), Stable Diffusion Esser et al. (2024), and Cogview 3 Zheng et al. (2024) have made significant strides in generating unique, realistic, and high-quality images from textual descriptions Gozalo-Brizuela & Garrido-Merchan (2023). Despite these achievements, there is still considerable room for improvement, especially in producing higher-resolution images that more accurately capture the nuances of the input text and in developing more intuitive and user-friendly interfaces Frolov et al. (2021). A persistent challenge is the difficulty these models face in understanding the subtle intentions behind user instructions, often resulting in a disconnect between user expectations and the generated outputs.

Manipulating input variables is inherently complex and challenging for non-expert users who lack formal prompt engineering training. The complexity of these variables can make it difficult to achieve the desired outcomes. Additionally, identical prompts often lead to varied image outputs in terms of content, layout, background, color, and perspective. This variability typically requires multiple attempts to generate an image that meets the user's expectations, making the process time-consuming.

To address this, we introduce the TDRI framework, designed to enhance the user experience by refining image outputs through iterative feedback. Unlike traditional models that rely heavily on prompt engineering, TDRI uses multi-turn interactions to better capture user objectives. By maintaining a continuous feedback loop, TDRI reduces the need for multiple trial-and-error attempts, simplifying the process and improving the quality and relevance of generated images. the quality and relevance of the resulting images. Our main contributions are:

- We investigate specialized human-machine interaction techniques tailored for interactive image generation, guiding users through a refined process that effectively captures and translates their intentions into visual outputs.

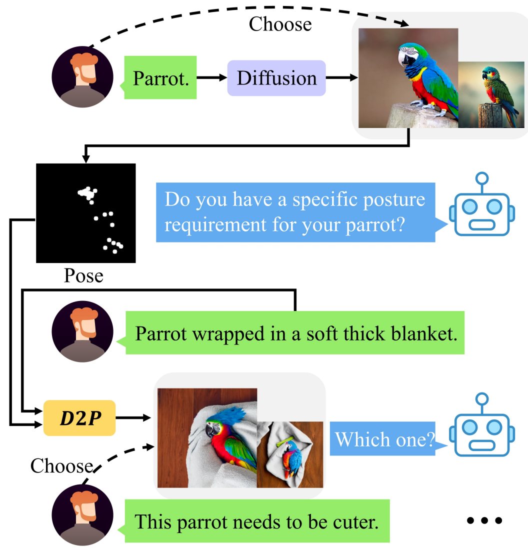


Figure 1: A multi-round dialogue interaction where the user refines the parrot's appearance using the Dialogue-to-Prompt (D2P) module (see Section 3.2.1). The system updates the image based on user feedback and pose constraints.

- We introduce the TDRI framework, a two-phase dialogue-based methodology for interactive image generation, which combines external user interactions with internal optimization processes to accurately render concepts into tangible images.
- We demonstrate the applicability of TDRI across various image generation tasks, highlighting its versatility and potential to revolutionize creative workflows by enabling rapid visualization and iteration of diverse concepts.

## 2 Related Work

### 2.1 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.

### 2.2 Ambiguity Resolution in Text-to-Image Generation

From visual annotations Endo (2023) and model evaluation benchmarks Lee et al. (2024) to auto-regressive 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.

### 2.3 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.

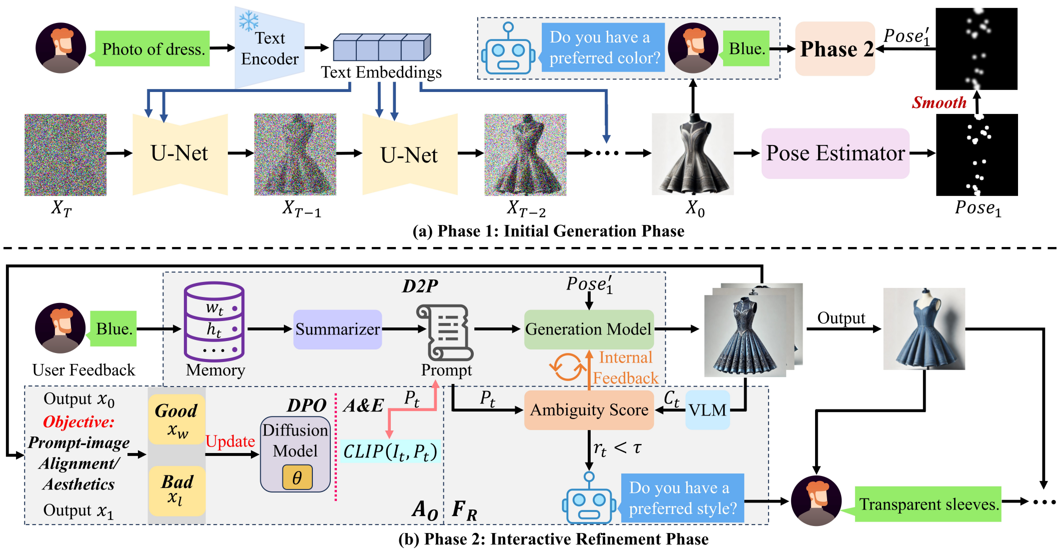


Figure 2: An overview of the two-phase framework TDRI. (a) In the Initial Generation Phase, the system processes user prompts via a U-Net-based diffusion model, generating base images with pose constraints. (b) In the Interactive Refinement Phase, user feedback is integrated to iteratively refine the image through dialogue-to-prompt generation, ambiguity scoring, and adaptive optimization.

## 3 Proposed Method

We propose a two-phase framework for image generation in multi-turn dialogues: the Initial Generation Phase, where the system processes the user's initial prompt ($w_1$) to generate an image ($I_1$) and extract pose ($\text{pose}_1$) as a constraint, and the Interactive Refinement Phase, where three modules—Dialogue-to-Prompt ($D2P$), Feedback-Reflection ($F_R$), and Adaptive Optimization ($A_O$)—iteratively refine the image based on user feedback to ensure comprehensive prompt representation.

### 3.1 Initial Generation Phase

The Initial Generation Phase initializes the image generation by processing the user input prompt $w_1$. The system generates a base image $I_1$ using a prompt-conditioned generative model $G(\cdot)$: $I_1 = G(w_1)$, where $I_1$ is the initial image generated based on prompt $w_1$. Subsequently, a pose estimator $\mathcal{P}(\cdot)$ extracts the pose $\text{pose}_1$ from $I_1$, represented by keypoint coordinates $\{(x_i, y_i)\}_{i=1}^{K}$ for $K$ keypoints: $\text{pose}_1 = \mathcal{P}(I_1)$. The extracted pose $\text{pose}_1$ acts as a structural constraint for subsequent iterations. A Gaussian smoothing function $\mathcal{S}(\cdot)$ is applied to refine $\text{pose}_1$, expanding its influence: $\text{pose}'_1 = \mathcal{S}(\text{pose}_1)$. This refined pose $\text{pose}'_1$ is used as a guiding feature in future image generation rounds, maintaining core structural integrity while allowing flexibility in user-directed updates.

### 3.2 Interactive Refinement Phase

#### 3.2.1 Dialogue-to-Prompt Module ($D2P$)

The *Dialogue-to-Prompt Module* ($D2P$) formulates the prompt $P_t$ at each timestep $t$ by integrating the dialogue history $h_t$ and the latest user input $w_t$. The dialogue history is defined as:

$$h_t = \{(w_1, r_1), (w_2, r_2), \ldots, (w_{t-1}, r_{t-1})\}, \tag{1}$$

where $w_i$ and $r_i$ represent the user input and system response at step $i$, respectively. The Summarizer $M_S$ synthesizes $h_t$ and $w_t$ to generate $P_t$:

$$\begin{aligned} P_t &= M_S(h_t, w_t) \\ &= g_{\text{sum}}\left(\sum_{i=1}^{t-1} \lambda_i \phi(w_i) + \mu_i \psi(r_i), \phi(w_t)\right), \end{aligned} \tag{2}$$

where $\lambda_i$, $\mu_i$ are weighting coefficients, $\phi(\cdot)$, $\psi(\cdot)$ are embedding functions mapping inputs to high-dimensional feature spaces, and $g_{\text{sum}}$ denotes the summarization operation. This aggregation ensures that $P_t$ encapsulates both historical context and current user intent, optimizing it for image generation. Subsequently, the Generation Model $M_G$ utilizes $P_t$ to produce the image $I_t$, conditioned on the initial pose $\text{pose}'_1$ and accumulated context $\mathcal{C}_{t-1}$:

$$I_t = M_G(P_t \mid \text{pose}'_1, \mathcal{C}_{t-1}), \tag{3}$$

where $\mathcal{C}_{t-1}$ aggregates contextual information from prior iterations.

#### 3.2.2 Feedback-Reflection Module ($F_R$)

The *Feedback-Reflection Module* ($F_R$) evaluates the generated image $I_t$ by extracting a set of descriptive features or captions, $C_t = \{C_t^1, C_t^2, \ldots, C_t^N\}$, where each $C_t^i$ represents a distinct characteristic of the image. In our implementation, the extraction function $f_E$ is handled by a vision-language model (VLM), specifically Qwen-VL (Bai et al., 2023). We incorporate specific prompt templates to guide the VLM in assessing the completeness of the generated image:

$$C_t = f_E(I_t) = \left\{C_t^i \mid i = 1, 2, \ldots, N\right\}, \tag{4}$$

where $f_E$ maps the image $I_t$ to a structured description $C_t$.

To evaluate the consistency between $P_t$ and $C_t$, a similarity measure $\sigma(P_t, C_t)$ is used to compute the discrepancy between the prompt and generated image. This results in an ambiguity score $r_t$: $r_t = 1 - \sigma(P_t, C_t)$, where $r_t \in [0, 1]$ indicates the level of mismatch. The function $\sigma(P_t, C_t)$ is defined as:

$$\sigma(P_t, C_t) = \frac{\sum_{i=1}^{N} \nu_i \kappa(P_t^i, C_t^i)}{\sum_{i=1}^{N} \nu_i}, \tag{5}$$

where $\kappa(P_t^i, C_t^i)$ represents a similarity function between the $i$-th component of the prompt and the corresponding feature in the generated image, and $\nu_i$ denotes a weight assigned to each feature's importance in the evaluation.

When the ambiguity score $r_t$ exceeds a threshold $\tau$, the system seeks further user input to refine the prompt. This process generates a clarification query $q_{t+1}$, which is formulated as:

$$q_{t+1} = f_{\text{clarify}}(P_t, C_t, r_t), \tag{6}$$

where $f_{\text{clarify}}$ is a function that analyzes the prompt $P_t$, image captions $C_t$, and the ambiguity score $r_t$ to determine the most relevant aspect of the ambiguity.

#### 3.2.3 Adaptive Optimization Module ($A_O$)

The *Adaptive Optimization Module* ($A_O$) integrates *Direct Preference Optimization* ($DPO$) and *Attend-and-Excite* ($A\&E$) to ensure alignment between generated images and user preferences while maintaining prompt fidelity.

**Direct Preference Optimization** ($DPO$) leverages user preference pairs $\mathcal{P} = \{(x_w, x_l)\}$, where $x_w$ is the preferred image and $x_l$ is the less preferred one. The goal is to maximize the likelihood of generating $x_w$ over $x_l$:

$$\mathcal{L}_{DPO}(\theta) = \mathbb{E}_{(x_w,x_l)\sim\mathcal{P}}\left[\log\frac{\pi_\theta(x_w \mid s)}{\pi_\theta(x_l \mid s)}\right]. \tag{7}$$

**Attend-and-Excite** ($A\&E$) ensures that all key elements from the input prompt $P_t$ are adequately represented in the image $I_t$. The misalignment loss is defined as:

$$L = 1 - \mathrm{Sim}(I_t, P_t), \tag{8}$$

where the similarity score $\mathrm{Sim}(I_t, P_t)$ measures the alignment between the image and the prompt. The gradient $\Delta P_t = \nabla_{P_t} L$ is computed to identify under-represented elements.

During training, ControlNet is tuned using the combined loss function:

$$\mathcal{L}_{A_O}(\theta) = \mathcal{L}_{DPO}(\theta) + \lambda\mathcal{L}_{A\&E}(\theta), \tag{9}$$

where $\lambda$ controls the balance between preference alignment and prompt fidelity.

# 4 Experiment

We evaluated the performance of the TDRI framework in two scenarios: fashion product creation and general image generation. Each scenario presents unique requirements. We first focused on fashion product creation due to the availability of a larger dataset, allowing us to capture fine-grained intent and user preferences. After demonstrating the model's success in this domain, we extended the framework to the general image generation task, where the focus shifted towards satisfying broader user intent.

## 4.1 Task 1: Fashion Product Creation

### 4.1.1 Setting

Fashion product creation poses greater challenges than general image generation due to higher demands for quality and diversity. Our Agent system requires advanced reasoning and multimodal understanding, supported by ChatGPT-4 for reasoning tasks. For image generation, we used the SD-XL 1.0 model, fine-tuned with the DeepFashion dataset (Liu et al., 2016) for clothing types and attributes. The LoRA (Hu et al., 2021) method was applied for fine-tuning on four Nvidia A6000 GPUs, resulting in more consistent outputs.

To provide a personalized experience, we trained multiple models with different ethnic data, allowing users to choose according to preferences. Using Direct Preference Optimization (DPO), model parameters were updated after every 40 user feedback instances, repeated three times, with the DDIM sampler for image generation.

### 4.1.2 Result Analysis

Figure 3 shows the outputs of six models optimized based on user selections and interaction history. All models generated fashion products from the same prompt using identical seeds, resulting in subtle variations. We collected feedback from six users and optimized the models with DPO, revealing distinct latent space characteristics under the same random seed. User evaluations showed significant performance improvements, with most testers preferring the DPO-optimized outputs (Figure 4).

## 4.2 Task 2: General Image Generation

### 4.2.1 Setting

In this task, the Summarizer generates prompts by aggregating the user's input, which are then used to create images. These images are captioned by Qwen-VL (Bai et al., 2023), a Vision-Language Model, across seven aspects: 'Content', 'Style', 'Background', 'Size', 'Color', 'Perspective', and 'Others'. We compare the CLIP similarity scores between the current generated image and each


Figure 3: This image presents a variety of fashion models and outfits, segmented by user preferences, showcasing styles from elegant dresses to casual and professional jackets, modeled by individuals of diverse ethnicities.

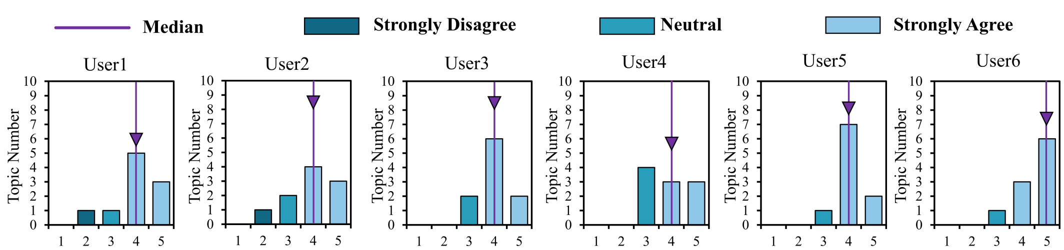


Figure 4: Human Voting for Statement: Direct Preference Optimization can improve generation results.

caption to identify ambiguous aspects. One of the three lowest-scoring aspects is randomly selected for questioning, and the user can choose to respond. In human-in-the-loop image generation, a target reference image is set, and user feedback is provided after each generation, with similarity to the target image used to assess effectiveness.

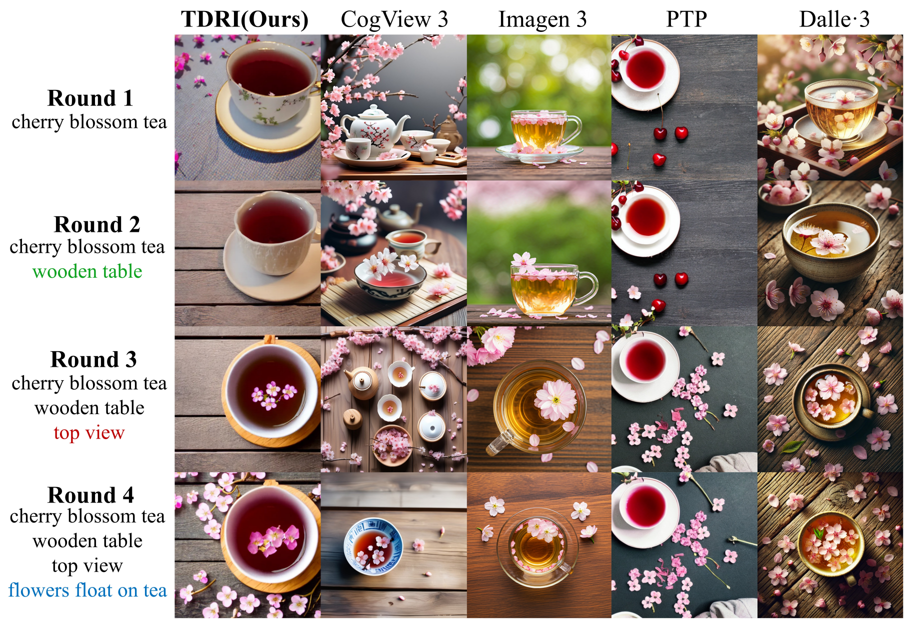


Figure 5: Comparison of cherry blossom tea images generated across four rounds by various models.

Table 1: Evaluations of prompt-intent alignment, image-intent alignment, and human voting across various methodologies and integrations

| Methods | Prompt-Intent Alignment | | Image-Intent Alignment | | Human Voting |
|---|---|---|---|---|---|
| | T2I CLIPscore | T2I BLIPscore | I2I CLIPscore | I2I BLIPscore | |
| GPT-3.5 augmentation | 0.154 | 0.146 | 0.623 | 0.634 | 5% |
| GPT-4 augmentation | 0.162 | 0.151 | 0.647 | 0.638 | 6.2% |
| LLaMA-2 augmentation | 0.116 | 0.133 | 0.591 | 0.570 | 6.1% |
| Yi-34B augmentation | 0.103 | 0.124 | 0.586 | 0.562 | 4.3% |
| TDRI-Reflection | 0.281 | 0.285 | 0.753 | 0.767 | 25.8% |
| TDRI-Reflection + ImageReward RL Xu et al. (2024) | 0.297 | 0.284 | 0.786 | 0.776 | 26.5% |
| **TDRI (Ours)** | **0.338** | **0.336** | **0.812** | **0.833** | **33.6%** |

#### 4.2.2 Data Collection

We curated 496 high-quality image-text pairs from the ImageReward dataset (Xu et al., 2024), focusing on samples with strong alignment to prompts. By removing abstract or overly complex prompts, as very long prompts tend to reduce accuracy and fail to clearly reflect the user's intent, we included people, animals, scenes, and artworks. Over 2000 user-generated prompts were used, with some images containing content not explicitly mentioned in the prompts. Each sample underwent at least four dialogue rounds for generation.

#### 4.2.3 Baseline Setup

To demonstrate the effectiveness of our Reflective Human-Machine Co-adaptation Strategy in uncovering users' intentions, we established several baselines. One method to resolve ambiguity in prompts is using Large Language Models (LLMs) to rewrite them. We employed various LLMs, including **ChatGPT-3.5**, **ChatGPT-4** (Achiam et al., 2023), **LLaMA-2** (Touvron et al., 2023), and **Yi-34B** (AI et al., 2024). Table **??** shows the alignment between generated prompts, target images, and output images. A subjective visual evaluation (Human Voting) was used to select the image closest to the target. All experiments were conducted on four Nvidia A6000 GPUs using the SD-1.4 model with the DDIM sampler.

Table 2: Ablation study of multi-dialog models across different rounds and metrics

| Multi-dialog | SD-1.4 | | SD-1.5 | | DALL-E 3 | | MetaGPT | | PTP | | CogView 3 | | Imagen 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP | BLIP | CLIP | BLIP | CLIP | BLIP | CLIP | BLIP | CLIP | BLIP | CLIP | BLIP | CLIP | BLIP |
| Round 1 | 0.728 | 0.703 | 0.723 | 0.699 | 0.651 | 0.674 | 0.646 | 0.672 | 0.661 | 0.681 | 0.643 | 0.664 | 0.671 | 0.691 |
| Round 2 | 0.759 | 0.738 | 0.746 | 0.725 | 0.675 | 0.690 | 0.671 | 0.691 | 0.682 | 0.700 | 0.667 | 0.679 | 0.696 | 0.712 |
| Round 3 | 0.776 | 0.764 | 0.773 | 0.784 | 0.691 | 0.718 | 0.689 | 0.711 | 0.701 | 0.716 | 0.684 | 0.696 | 0.727 | 0.732 |
| Round 4 | 0.804 | 0.824 | 0.790 | 0.811 | 0.743 | 0.736 | 0.726 | 0.742 | 0.712 | 0.726 | 0.705 | 0.717 | 0.751 | 0.742 |

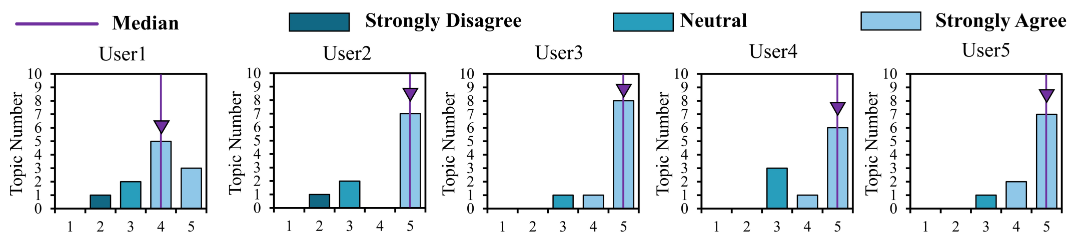


Figure 6: Human Voting for Statement: Multi-turn dialogues can approximate the user's potential intents.

Table 3: Attend-and-excite usage frequency and T2I similarity at different thresholds

| **Attend-and-Excite Threshold** | **0.80** | **0.75** | **0.73** | **0.70** | **0.68** | **0.66** |
|---|---|---|---|---|---|---|
| **Frequency of Usage** | 0 | 8.7 % | 31.3 % | 51.6 % | 72.5 % | 95.8 % |
| **T2I Similarity Improvement** | 0 | 0.23 % | 1.87 % | 2.36 % | 2.67 % | 1.3 % |

#### 4.2.4 RESULT ANALYSIS

**Visual and Quantitative Results** The visual results in Figure 5 demonstrate our reflective human-machine co-adaptation strategy. As user feedback refines through multiple dialogue turns, the generated images progressively align with the target images, showcasing our model's superior ability to adapt to user instructions. Tables **??** and **??** present experiments on our collected dataset. Table **??** compares the effectiveness of LLM augmentation for inferring user intent and evaluates the performance of our multi-dialog approach (TDRI-Reflection) using SD-1.4 and Qwen-VL. We also compare our TDRI method with a reinforcement learning approach using ImageReward (Xu et al., 2024) feedback. 'Intent' refers to target images, and similarity scores between prompts and images are measured using CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). User votes indicate which method best matched the target images, with our approach showing optimal performance. Table **??** highlights the effectiveness of HM-Reflection in resolving ambiguity across models, with image similarity improving over multiple dialog rounds.

**User Feedback** Figure 6 collects the approval ratings from five testers. In these dialogues, 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 closely aligned with user intent.

**Attend-and-Excite Performance** We also conducted independent experiments on Algorithm **??** (Attend-and-Excite) using the dataset from Task 2, with details of the algorithm provided in Appendix **??**. As shown in Table 3, the usage frequency of Attend-and-Excite varies with different thresholds $k$. At $k = 0.72$ and $k = 0.7$, the usage frequencies were 31.1% and 51.1%, respectively, with CLIP score increases of 1.8% and 2.3%, demonstrating that these settings improve image-text alignment.

**Embedding Refinement by Round** The t-SNE visualization in Figure 7 highlights how embeddings evolve across three interaction rounds. With each round of feedback, the embedding distribution becomes increasingly compact. It indicates that the model progressively refines its under-

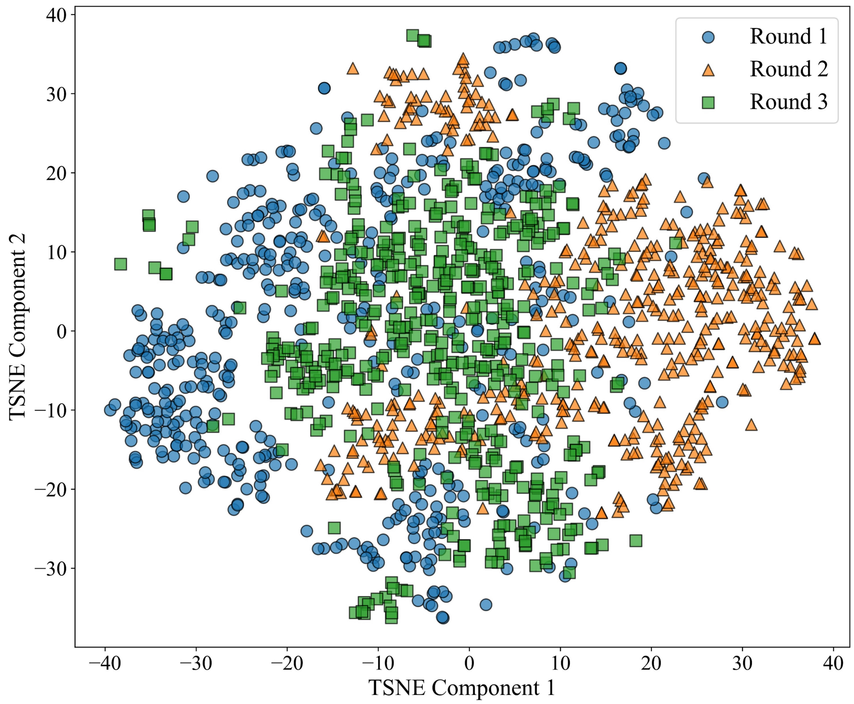


Figure 7: t-SNE visualization of embeddings across three interaction rounds.

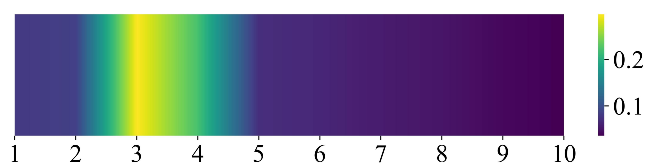


Figure 8: Heatmap showing user perception of the model's ability to capture intent across different dialogue rounds. The intensity peaks around 3 rounds.

standing of user intent, as seen by the tighter clustering of similar samples and reduced overlap between rounds. These improvements demonstrate the model's ability to capture user preferences more effectively through iterative optimization (refer to Tables 1 and 2).

**User Perception of Intent Capture** Figure 8 presents a heatmap illustrating user perception of the model's ability to capture intent across different dialogue rounds. The intensity peaks around the third round, indicating that users felt the model most accurately understood their intent at this stage. This suggests that by the third interaction, the model has significantly improved its comprehension of user preferences, and subsequent rounds provide only marginal gains in refining user intent.

**User Interaction Distribution by Round** The distribution of user interactions across dialogue rounds is shown in Figure 9. The majority of users required around five rounds to refine their image generation, with the highest proportion (21.1%) achieving their desired results by the fifth round. This suggests that the TDRI framework effectively captures user preferences within a relatively small number of interactions, with diminishing returns in later rounds as fewer users required additional feedback beyond round five.

## 5 Conclusion

In this study, we explored advanced image generation techniques combined with human-machine interaction to enhance personalization and improve visual outcomes in general image generation and

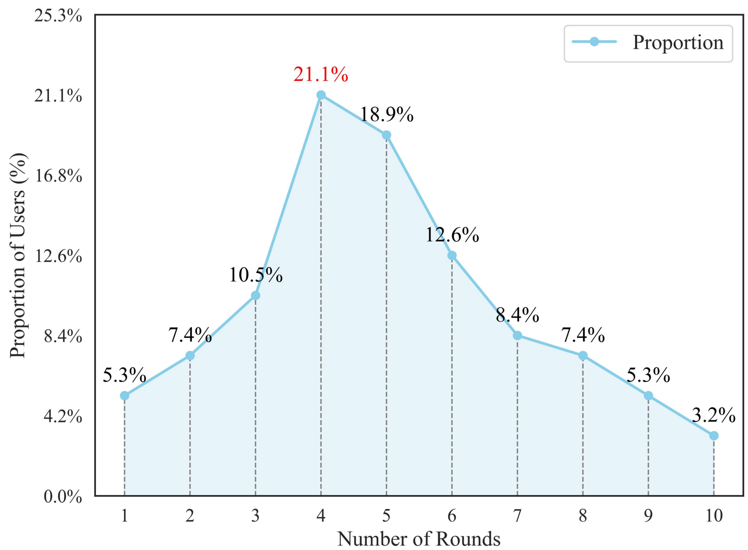


Figure 9: Proportion of users across dialogue rounds in the TDRI framework peaks at 5 rounds (21.1%), indicating most users refined their image generation within 5 interactions.

fashion product creation. Our TDRI framework effectively refined user intentions through dialogue-driven interactions and feedback reflection, progressively aligning outputs with user preferences. Its adaptability allowed it to handle diverse tasks, showing potential across various domains. Future work will focus on integrating granular feedback mechanisms and leveraging AI advancements to further optimize the process, extending the framework's versatility across creative and industrial applications.

# 6 LIMITATIONS

While TDRI offers significant improvements, it has certain limitations. The model may struggle to accurately translate complex, multi-level prompts into images due to the VL model's difficulty in capturing fine-grained details, leading to inaccurate captions. Additionally, cross-modal transfer errors can obscure user intent, reducing communication efficiency. The method is also computationally intensive and time-consuming, posing challenges for users with less powerful hardware. Future work should focus on enhancing efficiency and expanding the system's ability to generalize across diverse inputs to improve real-world usability.

# REFERENCES

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

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. Yi: Open foundation models by 01.ai, 2024.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *DALL-E 3*, 2023. OpenAI.

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

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

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

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

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

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

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

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

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

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

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

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

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 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)*, pp. 14367–14388, 2023.

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

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

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

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

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

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

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

Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion, 2024. URL https://arxiv.org/abs/2403.05121.

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

# A 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.

# B Guidelines for Human Annotation

## B.1 Objective

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

## B.2 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.

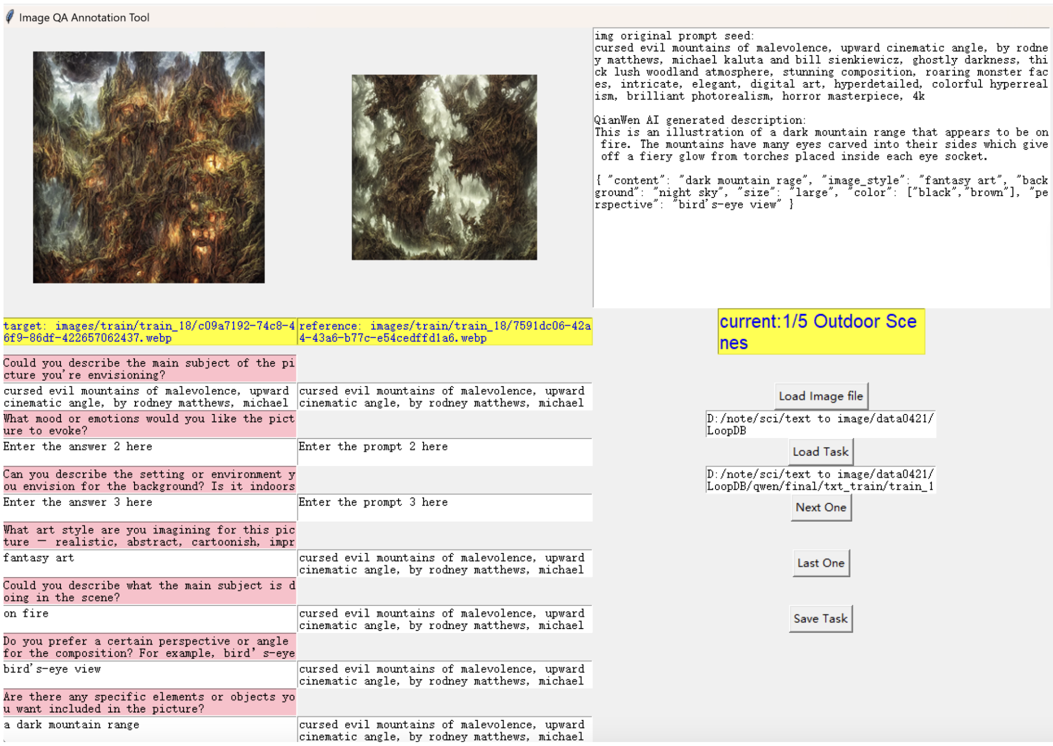


Figure 10: Screenshot of the Q&A software annotation interface

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.

### B.3 Usage 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.

For help, access the 'Help' section or contact the project manager at [contact information].

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

## C Attend-and-Excite

The Attend-and-Excite (A&E) algorithm is designed to iteratively refine an image generation process based on a given prompt. The process works by calculating the similarity between the current image $I_t$ and the user prompt $P_t$ using a CLIP-based similarity score. If the similarity exceeds a predefined threshold $k$, the process terminates. Otherwise, the algorithm calculates an objective to guide improvements and identifies the most significant tokens responsible for discrepancies between the image and the prompt. These tokens are appended to a list, and the image is regenerated with updated parameters, ensuring that the generated content aligns more closely with the user's intent through each iteration. This method allows for precise adjustments based on user feedback.

**Algorithm 1** Attend-and-Excite

**Require:** Image $I_t$, Prompt $P_t$

1: Initialize $token\_list \leftarrow \emptyset$, Iteration Number $N$, Threshold $k$
2: **for** $n = 1$ to $N$ **do**
3: Computing the Similarity of $I_t$ and $P_t$: $Sim \leftarrow$ 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 $l$: $\Delta P_t$
9: Locate peak value of $\Delta P_t$ 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 $I_t$

Table 4: Comparison of image editing and from scratch generation

| Method | Consistency Score | User Satisfaction (%) | Time Taken (minutes) |
|---|---|---|---|
| From Scratch | 0.75 | 78% | 12 |
| Image Editing | 0.88 | 90% | 9 |

Table 5: Comparison of simple vs. complex prompts

| Prompt Type | Generation Success Rate (%) | Average CLIP Score | Human Voting (%) |
|---|---|---|---|
| Simple Prompts | 92% | 0.85 | 87% |
| Complex Prompts | 65% | 0.60 | 62% |

Table 6: Generalized model vs. sample-specific D3PO

| D3PO Training Method | User Satisfaction (%) | Time to Convergence (iterations) | CLIP Score |
|---|---|---|---|
| Generalized Model | 83% | 5 | 0.77 |
| Sample-Specific Model | 90% | 8 | 0.85 |

# D Experimental Analysis

## D.1 Image Editing vs. From Scratch Generation

As shown in Table 4, Image Editing significantly outperforms the From Scratch method in terms of consistency (0.88 vs. 0.75) and user satisfaction (90% vs. 78%). Additionally, Image Editing requires less time (9 minutes vs. 12 minutes). This indicates that editing an existing image rather than generating from scratch leads to a more refined and efficient process, aligning closely with user expectations.

## D.2 Complex Prompt Exclusion Justification

In Table 5, the generation success rate is much higher for simple prompts (92%) compared to complex prompts (65%). The average CLIP score and human voting results also demonstrate that simple prompts are more effective in generating images that align with user intent. The drop in performance with complex prompts (CLIP score of 0.60 vs. 0.85 for simple prompts) supports the decision to focus on excluding overly complex prompts for more consistent results.

Table 7: Effect of interaction turns on image quality, satisfaction, and time

| Turns | Satisfaction (%) | CLIP Score | Time (min) |
|---|---|---|---|
| 2 | 70% | 0.72 | 6 |
| 4 | 85% | 0.78 | 9 |
| 6 | 87% | 0.80 | 11 |
| 8 | 88% | 0.81 | 12 |

Table 8: Performance comparison of lightweight models

| Model Size | User Satisfaction (%) | CLIP Score | Computation Time (minutes) |
|---|---|---|---|
| 7B | 90% | 0.85 | 15 |
| 5B | 85% | 0.82 | 10 |
| 3B | 78% | 0.77 | 6 |

### D.3 Generalized vs. Sample-Specific D3PO

Table 6 highlights that while the sample-specific model achieves higher user satisfaction (90%) and a better CLIP score (0.85), it requires more iterations to converge (8 vs. 5 for the generalized model). This suggests that sample-specific tuning can yield higher-quality results, though at the cost of additional computation time and iterations.

### D.4 Ablation Study on Interaction Turns

From the results in Table 7, we observe that increasing the number of interaction turns improves both user satisfaction and image quality. Specifically, satisfaction rises from 70% at 2 turns to 88% at 8 turns, while the CLIP score increases from 0.72 to 0.81. However, the marginal improvement between 6 and 8 turns is small, suggesting diminishing returns beyond 6 turns, and with a noticeable increase in time taken (from 11 to 12 minutes).

### D.5 Lightweight Models Comparison

Table 8 shows that the 7B model achieves the highest user satisfaction (90%) and CLIP score (0.85) but also takes the longest computation time (15 minutes). Meanwhile, the 3B model is the fastest (6 minutes), but at the expense of lower user satisfaction (78%) and CLIP score (0.77). This indicates that while smaller models offer faster results, they compromise on image quality and user satisfaction, and a balance between performance and speed must be considered depending on the task.