DialogDraw: Image Generation and Editing System Based on Multi-Turn Dialogue

Shichao Ma*†, Xinfeng Zhang†, Zeng Zao‡, Bai Liu, Changjie Fan, Zhipeng Hu

Fuxi AI Lab, NetEase Inc.
mashichao@mail.ustc.edu.cn, {zhangxinfeng01, hzzhaozeng, hzliubai, fanchangjie, zphu}@corp.netease.com

Abstract

In recent years, diffusion modeling has shown great potential for image generation and editing. Beyond single-model approaches, various drawing workflows now exist to handle diverse drawing tasks. However, few solutions effectively identify user intentions through dialogue and progressively complete drawings. We introduce DialogDraw, which facilitates image generation and editing through continuous dialogue interaction. DialogDraw enables users to create and refine drawings using natural language and integrates with numerous open-source drawing workflows and models. The system accurately recognizes intentions and extracts user inputs via parameterization, adapts to various drawing function parameters, and provides an intuitive interaction mode. It effectively executes user instructions, supports dozens of image generation and editing methods, and offers robust scalability. Moreover, we employ SFT and RLHF to iterate the **I**ntention **R**ecognition and **P**arameter **E**xtraction Model (IRPEM). To evaluate DialogDraw's functionality, we propose DrawnConvos, a dataset rich in drawing functions and command dialogue data collected from the open-source community. Our evaluation demonstrates that DialogDraw excels in command compliance, identifying and adapting to user drawing intentions, thereby proving the effectiveness of our method.

1 Introduction

Recently, significant advancements have been achieved in the field of image generation using diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2020; Podell et al. 2023). These large-scale text-to-image models can synthesize high-quality, diverse images from concise textual prompts. The adoption of various diffusion models for a wide array of applications, including image generation, editing, and creative work, is on the rise. Furthermore, the open-source community centered around diffusion models is thriving. Platforms like Civitai (Civitai 2022) and OpenArt (OpenArt 2022) are particularly noteworthy, where numerous expert users share models and workflows. This collaborative environment not only fosters the advancement of artistic creation but also unveils the vast potential of these technologies.

Simultaneously, there is a strong demand for image generation and continuous editing capabilities. Creating exceptional art often requires repeated adjustments. Currently, users typically need to manually download various models and workflows to iteratively refine their creations, a tedious process. The advent of conversational large language models (LLMs) has introduced a more streamlined and intuitive approach. For example, DALL-E 3 (Betker et al. 2023) demonstrates how multi-round dialogue systems facilitate concise and clear image generation and editing. This method has gained public acceptance and is driving research in multi-round dialogue-based drawing techniques.

Most current drawing workflows that incorporate Large Language Models (LLMs) primarily rewrite and expand the user's initial input before feeding it into the text model to generate images. This approach often fails to capture the user's intent, such as making subtle adjustments or converting styles. Works like InstructPix2Pix (Brooks, Holynski, and Efros 2023), InstructDiffusion (Geng et al. 2024), and DialogPaint (Wei et al. 2023) rely on instruction editing through dialogue but often depend on a single model, limiting their ability to understand instructions and achieve effective edits. Moreover, there is limited research on generating and editing images through multi-round dialogues. For instance, DialogPaint (Wei et al. 2023) can only edit the original image, while DialogGen (Huang et al. 2024) regenerates prompts for dialogues, compromising image consistency. User needs are diverse, as evidenced by open-source communities like OpenArt (OpenArt 2022) and Civitai (Civitai 2022). Many existing instruction editing methods and single-model solutions struggle to meet these varied needs.

Therefore, continuous development and expansion of pipelines are crucial. Currently, there are hundreds of mature pipelines available on OpenArt (OpenArt 2022). The challenge is integrating these pipelines and models with LLM methods to accurately transform natural language descriptions into clear intents, accommodating diverse inputs to create rich and varied artworks. In this paper, we have developed a multi-turn dialogue-based drawing generation and editing system called DialogDraw. This conversational system is designed to generate, understand, and continuously edit images. Our system primarily comprises the In-

*Interns in Fuxi AI Lab, NetEase Inc.
†These authors contributed equally.
‡Corresponding Author.

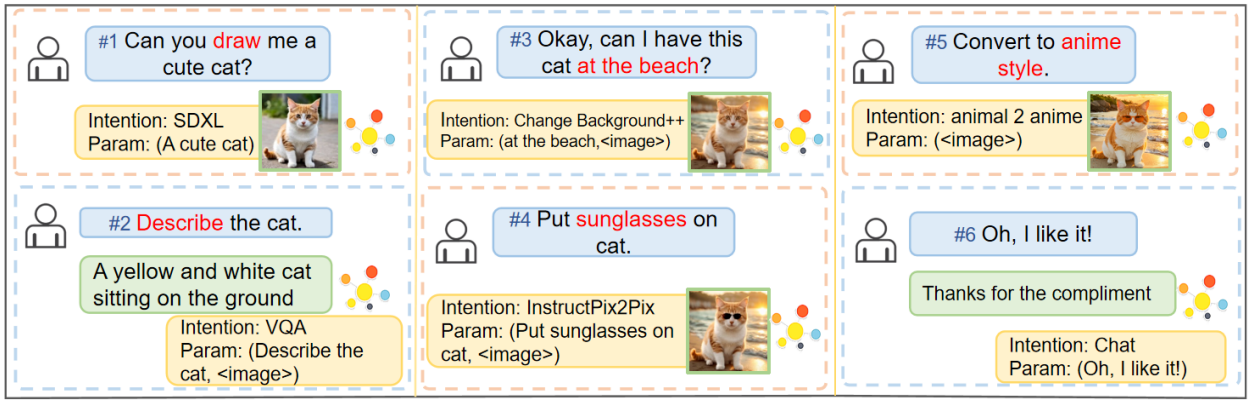

Figure 1: DialogDraw is based on multi-turn dialogue. It does intent recognition and parameter extraction (IRPE) for users' natural language inputs regarding image generation, editing, description, and natural dialogues. Blue is for Question, yellow for IntentionParam, and green for Response.

tent Recognition and Parameter Extraction Model (IRPEM) and drawing pipelines. Utilizing a multi-turn dialogue multimodal model (MLLM), we created a comprehensive simulated multi-modal dialogue dataset to train our model. By leveraging diverse multi-modal inputs from users and our trained multi-modal intent recognition model, we can accurately infer users' true intentions—whether they are making queries, generating images, or performing various editing operations and extracting the necessary input parameters for different models and pipelines.

Additionally, we introduce a dataset for training and evaluating multimodal generation and editing systems, named DrawnConvos. Utilizing ChatGPT (OpenAI 2022) and SDXL (Podell et al. 2023), we generated a dataset of multi-round drawings through a set of automated processes encompassing image generation, editing, and visual question answering (VQA). We also developed more comprehensive metrics to evaluate the adherence to multi-round drawing instructions and the accuracy of multi-round intent switching. In our evaluation, we compare our approach with recent work in the industry to demonstrate its effectiveness and practicality.

In summary, our contributions are primarily as follows:

- We propose DialogDraw, the first system to combine multiple pipelines for drawing and editing images in multi-turn dialogues, composed of the Intent Recognition and Parameter Extraction Model (IRPEM) and various drawing abilities.
- We create a new dataset named DrawnConvos, a dataset of multi-round dialogues including image generation and editing, which incorporates numerous open-source workflows and models. Using this dataset, we apply SFT and RLHF methods for IRPEM training.
- We introduce a benchmark for DialogDraw in multi-round drawing scenarios, including metrics like Multi-turn VQA Score, Multi-turn CLIP Similarity, and Instruct Edit Coherence. Extensive experiments validate the effectiveness of our approach.

2 Related Work

2.1 Diffusion Models and Community

Diffusion-based models (Ho, Jain, and Abbeel 2020; Song and Ermon 2020) have demonstrated outstanding performance in image generation, offering enhanced stability and controllability. These models employ a forward process that involves adding Gaussian noise to the input image, followed by an inverse process that generates high-quality images with intricate details and diversity from random Gaussian noise. The latent diffusion model (LDM) (Rombach et al. 2022) has been introduced to shift the diffusion process from the pixel space to the latent space, significantly improving both efficiency and image quality. There are already several diffusion model-based generation methods (GLIDE(Nichol et al. 2022), Imagen(Saharia et al. 2022), Stable Diffusion(Rombach et al. 2022), SDXL(Podell et al. 2023), Controlnet(Zhang, Rao, and Agrawala 2023), T2I-Adapter(Mou et al. 2024)) and editing (InstructPix2Pix (Brooks, Holynski, and Efros 2023), InstructDiffusion (Geng et al. 2024), etc.), many of which also have been integrated in open-source communities.

The open-sourcing of stable diffusion has led to the emergence of numerous derivative models and workflow communities. Notably, Stable Diffusion WebUI and ComfyUI have become popular frameworks for image generation. Additionally, communities like Civitai (Civitai 2022) and OpenArt (OpenArt 2022) have flourished, allowing users to customize and edit images based on the open-source platform.

2.2 Large Language Models

Large language models(Brown et al. 2020; Shuster et al. 2022; Wei et al. 2022; Zhang et al. 2019) have been widely studied in recent years, with the capability to chat with humans fluently. Models such as GPT-3 (Brown et al. 2020) can generate simulated data according to given samples, which is a convenient way to gather language data in a specific format and finetune other language models. Conversation-oriented language models like DialoGPT (Zhang et al. 2019), Meena (Adiwardana et al. 2020),

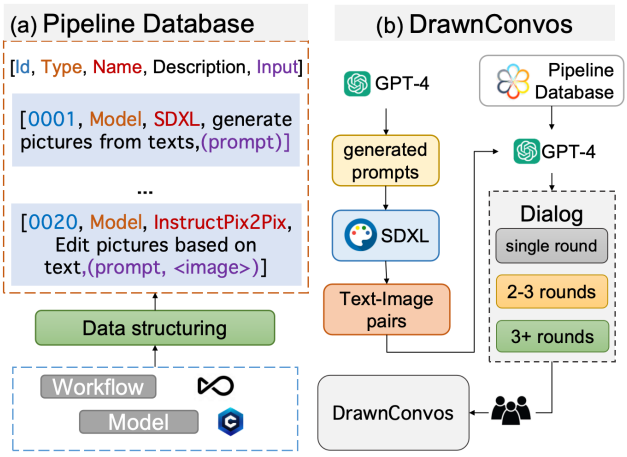

Figure 2: Overview of the Construction of DrawnConvos. (a) Construction of pipeline structured data (models and workflows). (b) Construction of DrawnConvos (including SFT, RLHF, and TEST)

BlenderBot (Roller et al. 2020), and ChatGPT (OpenAI 2022) have shown exceptional performance in various conversational tasks.

Multimodal large models (Chen et al. 2020; Ramesh et al. 2021; Jia et al. 2021; Singh et al. 2022; Alayrac et al. 2022; Li et al. 2022; Bai et al. 2023) integrate text, image, and audio data, enabling cross-modal information processing and understanding. CLIP (Radford et al. 2021) and DALL-E (Ramesh et al. 2021) achieved significant advancements in image generation and cross-modal retrieval by training on large-scale image and text data. Recent multimodal models like FLAVA (Singh et al. 2022), MUM (Jia et al. 2021), Qwen-VL(Bai et al. 2023) and Hunyuan (Li et al. 2024) exhibit excellent performance across more modalities and tasks, pushing the boundaries of artificial intelligence in handling complex information.

2.3 Dialog-based for Drawing

Research on generating and editing images through multi-turn dialogue has garnered increasing attention in recent years. Early works (Chen et al. 2018) explored modifying image attributes via natural language instructions and dialogue interactions. Subsequently, a GAN-based model utilized sequential attention mechanisms to edit images based on conversational inputs (Cheng et al. 2020), laying the groundwork for interactive image editing tools.

Advancements such as ChatEdit (Cui et al. 2023), DialogPaint (Wei et al. 2023), and DialogGen (Huang et al. 2024) have further propelled the field. ChatEdit focuses on facial image editing through dialogue and includes a constructed dataset for its studies. DialogPaint bridges conversational interactions with image editing, allowing users to modify images through natural dialogue. DialogGen, a multi-modal interactive dialogue system, addresses multi-turn, multi-modal image generation tasks, showcasing the potential of cross-modal interaction in image generation.

3 Methodology

The core of DialogDraw encompasses the structured data construction of the drawing workflows and models, the dialogue data construction, and the development of image generation and editing based on continuous dialogue.

3.1 Construction of DrawnConvos

First, we need to obtain the drawing pipelines (including workflows and models), their corresponding functional descriptions, and input parameters. We collected approximately 20 pipelines from OpenArt (OpenArt 2022) and Civitai (Civitai 2022), classifying them based on the number of downloads and varying functions. This classification ensures broad coverage of different drawing functionalities and promotes higher user engagement.

Fig. 2(a) illustrates how to build structured data for each pipeline. Initially, we assign each drawing pipeline an *Id*, *Name*, and *Type*. For the *Description*, we use GPT-4 (Achiam et al. 2023) to expand the original title based on the title of each pipeline and the existing page information. Finally, we define the *Input* parameters for each pipeline according to the downloaded pipeline. For instance, the "animal 2 anima" pipeline requires 1 input parameter ($< image >$), while the "InstructPix2Pix" pipeline requires 2 input parameters ($prompt, < image >$). At present, In our skill pipeline library, aside from the 20 mentioned, there are two more for non-image tasks: VQA for describing images, and Chat for conversing with users.

The construction of our dialogue dataset is primarily depicted in Fig. 2(b). Initially, following specific guidelines, we utilized GPT-4 to generate 500 prompts covering various categories such as people, animals, and landscapes. Then, using SDXL, we created corresponding images for these prompts, yielding 500 *(prompt, image)* pairs. Subsequently, based on these image-text pairs, we generated dialogues with GPT-4, crafting 3,000 single-turn, 6,000 two-to-three-turn dialogues, and 1,000 dialogues with more than three turns, totaling 10,000 dialogues. It should be noted that each multi-turn dialogue was constructed based on a single *(prompt, image)* pair.

This dataset is named "DrawnConvos." Our approach to dataset construction differs from others in that our model's primary output includes pipeline names and their corresponding parameters; image generation and editing occur within these pipelines. We then randomly divided DrawnConvos into DrawnConvos$_{(SFT)}$, DrawnConvos$_{(HF)}$, and DrawnConvos$_{(TEST)}$ in a 6:3:1 ratio.

DrawnConvos$_{(SFT)}$ In DrawnConvos$_{(SFT)}$, the distribution of dialogue turns still roughly approximates 30% single-turn dialogues, 60% 2-3 turn dialogues, and 10% dialogues exceeding three turns. It is used for Supervised Fine-Tuning (SFT) as a dataset for intent recognition and parameter identification to train IRPEM$_{(SFT)}$.

DrawnConvos$_{(RLHF)}$ In the second phase, IRPEM$_{(SFT)}$ is prompted with prompts x to generate pairs of answers $(y_1, y_2) \sim \pi_{SFT}(y|x)$. These pairs are then presented to human labelers who express a preference for one answer over

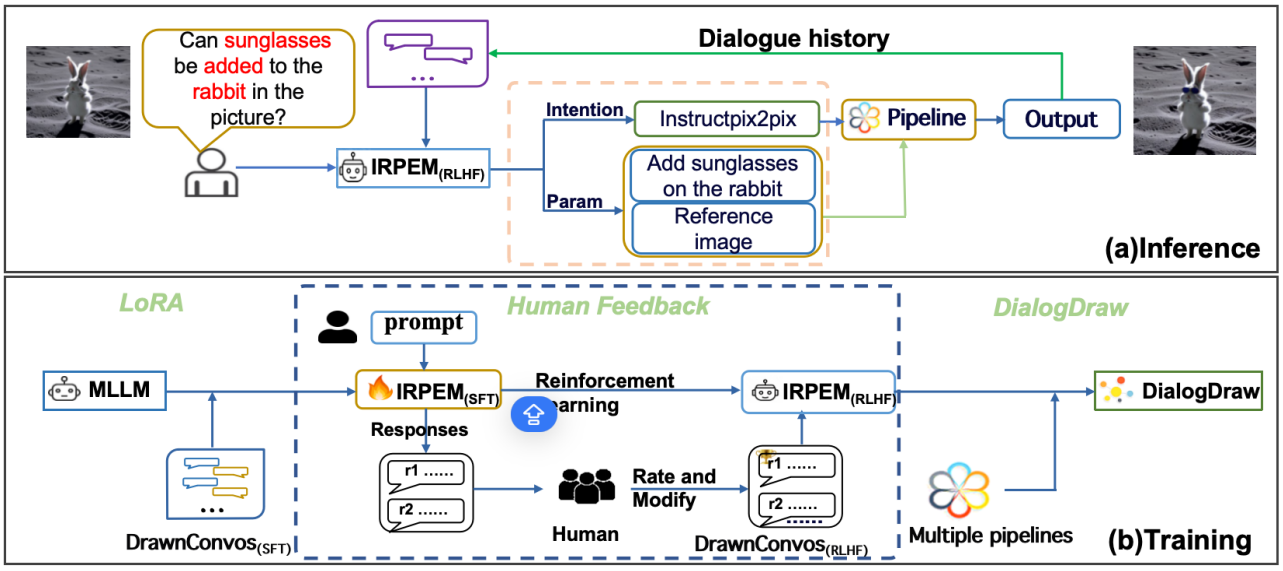

Figure 3: Overview of DialogDraw (a) DialogDraw's inference procedure: we extract user intent and parameters from dialogues and invoke the corresponding pipeline to edit the images. (b) DialogDraw's training procedure: using a two-stage training process and multiple pipelines to construct the system.

the other, denoted as $y_w \succ y_l$ for a given x, where y_w and y_l represent the preferred and less preferred completions, respectively, among the set (y_1, y_2). Notably, deviating from the previous strategy (Rafailov et al. 2024), when neither of the answer pairs (y_1, y_2) meets expectations, we modify them to better align with human preferences. Subsequently, we construct an offline dataset of preferences D consisting of elements $\{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}$ for $i = 1, \ldots, N$.

DrawnConvos$_{(TEST)}$ Similar to *DrawnConvos$_{(RLHF)}$*, *DrawnConvos$_{(TEST)}$* has also been annotated by humans. It is used to test the effectiveness of various models.

3.2 Construction of DialogDraw

As shown in Fig. 3(b), DialogDraw is a system for image generation and editing, composed of a model obtained through Reinforcement Learning from Human Feedback (Christiano et al. 2017; Stiennon et al. 2020; Bai et al. 2022) and various pipelines. Specifically, the process is divided into two steps: The first step involves supervised fine-tuning (Gunel et al. 2020; Yu et al. 2020) using the *DrawnConvos$_{(SFT)}$* dataset to train the model's basic intent and parameter recognition capabilities, resulting in the *IRPEM$_{(SFT)}$*. Utilizing *IRPEM$_{(SFT)}$*, the model's behavior is continuously adjusted through the Direct Proximal Optimization reinforcement learning framework (DPO) (Rafailov et al. 2024), culminating in the *IRPEM$_{(RLHF)}$*. Our system recognizes the user's intent and parameters by *IRPEM$_{(RLHF)}$* and then calls the appropriate pipeline and passes the appropriate parameters to edit and generate the image.

IRPEM$_{(SFT)}$ In the first phase of fine-tuning, the fine-tuning method is LoRA (Hu et al. 2021), and the dataset is *DrawnConvos$_{(SFT)}$*. The loss function is a cross-entropy loss function for predicting the next word:

$$L = -\sum_{t=1}^{T} \log\left(p(w_t|w_{1:t-1})\right) \tag{1}$$

where $p(w_t|w_{1:t-1})$ is the probability of word w_t given the previous words $w_{1:t-1}$, and T is the sequence length.

After the first step, we obtain the *IRPEM$_{(SFT)}$*. This model analyzes user questions to determine intent and corresponding parameters, providing a foundation for optimization.

IRPEM$_{(RLHF)}$ In this phase, the model builds upon the previous *IRPEM$_{(SFT)}$*, integrating *DrawnConvos$_{(RLHF)}$*, and continuously refines the model's behavior through the DPO reinforcement learning framework.

DrawnConvos$_{(RLHF)}$ can be represented as $D = \{(x_i^{(N)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^{N}$, where we can parameterize a reward model $r_\phi(x, y)$ and estimate the parameters via maximum likelihood. In reinforcement learning, we frame the problem as a binary classification task, where the negative log-likelihood loss is given by

$$\mathcal{L}_R(r_\phi, D) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right] \tag{2}$$

where σ is the logistic function.

In the DPO strategy, we have the probability of human preference data in terms of the optimal policy rather than the reward model, we can formulate a maximum likelihood objective for a parametrized policy π_θ. Analogous to Equation 2, our policy objective becomes:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D} \left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{3}$$

Model	Task				Quantitative Metrics			
	Image Generate	Image Edit	VQA	Chat	Multi-turn VQA Score ↑	Multi-turn CLIP Similarity ↑	Instruct Edit Coherence ↑	Human Score ↑
SEED-LLaMA-8B(Ge et al. 2023)	✓	×	✓	✓	0.7646	0.6719	0.7229	0.7117
SEED-LLaMA-14B(Ge et al. 2023)	✓	×	✓	✓	0.7781	0.6865	0.7369	0.7432
GPT-4(Achiam et al. 2023)	✓	×	✓	✓	**0.8523**	0.7412	0.8023	0.8218
DialogGen(Huang et al. 2024)	✓	×	✓	✓	0.8213	0.7217	0.7765	0.8081
DialogDraw(Ours)	✓	✓	✓	✓	0.8491	**0.7835**	**0.8196**	**0.8494**

Table 1: Comparison of quantitative analysis indicators for different models. ✓indicates that the model used this strategy during training, while × indicates that it did not. ↑ indicates that a higher score is better, and **bold** indicates the best results. The Instruct Edit Coherence score fully aligns with human evaluation.

By employing this method, we establish an implicit reward through a different parameter setup, with the optimal policy being π_θ. After this step, we obtain the ***IRPEM*$_{(RLHF)}$**.

DialogDraw The model for parameter and intent recognition, denoted as ***IRPEM*$_{(RLHF)}$**, is represented by M_D. Subsequently, the various pipelines are represented by P. Thereupon, our DialogDraw system can be expressed as the synthesis of M_D and P, represented mathematically as:

$$\text{DialogDraw} = M_D \oplus P \tag{4}$$

3.3 The Benchmark of DialogDraw

Multi-turn VQA Score. The VQA Score(Lin et al. 2024) serves as a robust metric for assessing the alignment between model-generated images and their corresponding prompt texts. Its approach is straightforward: it measures the generative likelihood of responses to simple questions in an end-to-end manner.

While effective for single-turn dialogues, this method falls short in evaluating the alignment across multiple dialogue turns. To overcome this limitation, we introduce the multi-turn VQA Score. The concept is as follows: for a dialogue represented as $(text_i, image_i)$, taking $i = 3$ for illustration, the calculation proceeds as follows: - The initial dialogue round uses the VQA Score for $(text_1, image_1)$. - The second round calculates the score for the integrated text $F(text_1, text_2)$ with $image_2$, where "F" signifies the concatenation of $text_1$ and $text_2$ using GPT-4 to form a comprehensive prompt for multi-turn dialogues. - The third round extends this approach to $((F(text_1, text_2, text_3), image_3))$.

The multi-turn VQA Score $S_{\text{M\_VQA}}$ is encapsulated by the formula:

$$S_{\text{M\_VQA}} = \frac{1}{n}\sum_{i=1}^{n} VQA(F(text_1, text_2, \ldots, text_i), image_i) \tag{5}$$

Here, n denotes the total number of dialogue turns, F is the function for integrating multi-turn text, and VQA is the scoring mechanism.

Multi-turn CLIP Similarity. To gauge the consistency of image generation and editing throughout multi-turn dialogues, we propose an additional metric utilizing the CLIP model (Radford et al. 2021). This metric, $S_{\text{M\_CLIP}}$, assesses the similarity between each pair of consecutive images:

$$S_{\text{M\_CLIP}} = \frac{2}{n(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1} \text{CLIP}\left(\text{image}_j, \text{image}_i\right) \tag{6}$$

This formula accounts for all possible pairs of images across the dialogue turns, with n being the total number of turns and CLIP the function that computes the similarity between two images.

Instruct Edit Coherence. It's important to clarify that this metric is predicated on the iterative editing of the same image. If a new image is generated in a particular round, the aforementioned formula would not be applicable.

We further refine our assessment by combining the multi-turn VQA Score and the CLIP Similarity, assigning weights q_1 and q_2 respectively. The weighting is adjusted based on the nature of the dialogue round: - For image editing rounds, the weights are balanced at $q_1 : q_2 = 0.5 : 0.5$. - For rounds involving new image generation, the weights are set to $q_1 : q_2 = 1 : 0$.

The final metric to measure the multi-turn dialogue's command understanding capability and the consistency of image generation and editing, Instruct Edit Coherence (IEC), is defined as:

$$IEC = q_1 \times S_{\text{M\_CLIP}} + q_2 \times S_{\text{M\_CLIP}} \tag{7}$$

It should be noted that our system includes "Chat" and "VQA" skills, designed for interactions not related to image generation or editing. In such cases, these instances are excluded from the scoring calculation.

4 Experiments

4.1 Experimental Setup

All our experiments are performed on four NVIDIA A100 GPUs using the PyTorch framework. During the training phase, we initialized our model with a pre-trained Qwen-VL (Bai et al. 2023) model. In the first phase, we trained the model for 50 epochs using ***DrawnConvos*$_{(SFT)}$** to obtain ***IRPEM*$_{(SFT)}$**. The second phase was built upon the first, where we further trained the model for another 50 epochs using ***DrawnConvos*$_{(RLHF)}$** to achieve ***IRPEM*$_{(RLHF)}$**. Both phases utilized the AdamW optimizer with weight decay set to 0.1 and 0.05, respectively. The initial learning rate for both stages is initialized as 1e-5.

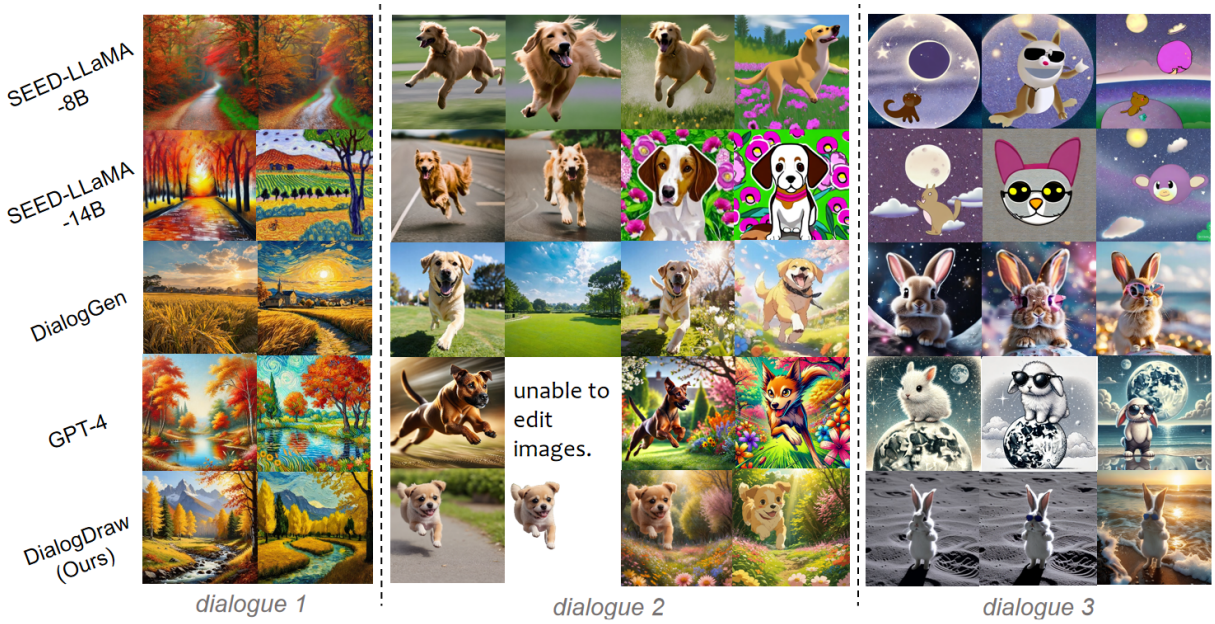

Figure 4: Visualization of Dialog Outputs. The above are three multi-round dialogues. In **dialogue1**, first, a painting of autumn scenery is wanted, then it's asked to be in Van Gogh's style. **Dialogue2** starts with a request for a running dog image, followed by asking to draw it, changing the background to a spring garden, and finally converting to anime style. For **dialogue3**, initially, a rabbit on the moon is requested, then sunglasses are added, and the background is changed to the beach.

4.2 Analysis and Comparisons

Table 1 shows the results of different models on the test set, and we can analyze their performance using the following evaluation metrics. In Fig. 4, we present images generated by different models.

Multi-turn VQA Score. In the Multi-turn VQA Score metric, as shown in Table 1, DialogDraw ranks second with a score of 0.8491, just behind GPT-4. This metric assesses the model's ability to transform text into images. DialogDraw utilizes SDXL for drawing and does not particularly pursue drawing capabilities, whereas GPT-4 excels in this area, hence its higher score. However, unlike the previous single-turn text-to-image generation, this metric considers the integration of previous text rounds, emphasizing the significance of contextual relationships—an area where DialogDraw excels. Consequently, DialogDraw scores higher than DialogGen and the other two models. In Fig. 4, DialogGen's lower score results from inaccuracies in understanding specific texts. For example, in dialogue 2, where the user intended to receive an image of a specific object, DialogGen incorrectly provides only the background without the object.

Multi-turn CLIP Similarity. DialogDraw secures the top position with a score of 0.7835, which is 5.71% higher than the second place. As shown in Fig. 4, other models such as DialogGen and seed-llama essentially generate new images for editing tasks rather than editing the original image; GPT-4 behaves similarly, and its inability to perform certain image editing instructions, like a cutout, leads to a lower score. DialogDraw, with its capability to directly edit the original image, achieves the highest score in this category, indicating the best consistency in image editing.

Instruct Edit Coherence. DialogDraw leads with a score of 0.8196. This score is a composite of the previous two metrics and reflects the model's understanding of instructions and the consistency of image generation and editing in multi-turn dialogues. DialogDraw's superior performance here is attributed to its excellent consistency in multi-turn image generation and editing; moreover, GPT-4 and DialogGen rank second and third, respectively. It is noteworthy that SEED-LLaMA performs poorly across these metrics.

Hunan Score. We invited users to rate image generation and editing in dialogues from different models. In the evaluation by users, two criteria are employed to assess the performance: the model's ability to comprehend instructions within multi-turn dialogues, and the consistency between image generation and editing. Both metrics are given equal weight. Randomly picking 100 dialogues, we received feedback from 50 users. As shown in Table 1, DialogDraw is favored over the Baseline, ranking higher in relevance and user preference. It scores 0.8350 for understanding dialogue commands, placing it just below GPT-4. For image consistency, it achieves a 0.8638 score. Overall, it tops the Human Score category with 0.8494, validating the Instruct Edit Coherence(IEC) metric.

4.3 Ablation Experiments

In this section, we have constructed a series of ablation studies to further analyze IRPEM and the intent and parameter recognition framework that we have proposed.

Model	Training Method		Quantitative Metrics		
	SFT	DPO	Intent Switch Accuracy ↑	Parameter Accuracy ↑	Structural Conformity ↑
Qwen-VL-zero-shot	×	×	0.702	0.754	0.738
w/ {Supervised Fine-Tuning }	×	✓	0.902	0.870	0.932
w/ {Direct Preference Optimization }	✓	×	0.874	0.857	0.865
$\text{IRPEM}_{\text{(RLHF)}}$	✓	✓	**0.924**	**0.892**	**0.953**

Table 2: Results of ablation study to analyze different components.

Model	Quantitative Metrics			
	Multi-turn VQA Score ↑	Multi-turn CLIP Similarity ↑	Instruct Edit Coherence ↑	Human Score ↑
Qwen-VL-zero-shot + Pipelines	0.7613	0.7317	0.7480	0.7542
w/ {Supervised Fine-Tuning } + Pipelines	0.8287	0.7707	0.8026	0.8228
w/ {Direct Preference Optimization } + Pipelines	0.8082	0.7584	0.7858	0.8041
$\text{IRPEM}_{\text{(RLHF)}}$ + Pipelines	**0.8491**	**0.7835**	**0.8196**	**0.8494**

Table 3: Ablation study on the effects of different strategies for multi-turn dialogue image generation and editing.

Quantitative Metrics. To present the results more intuitively, we have quantified the outputs of various models on $DrawnConvos_{(TEST)}$, focusing on three key metrics:

- **Intent Switch Accuracy (ISA).** ISA measures the match between the model's identified intents and the test set's standard intents:

$$\text{ISA} = \frac{1}{n}\sum_{i=1}^{n} 1_{\text{correct\_intent}_i} \tag{8}$$

  Here, ISA represents the Intent Switch Accuracy, n is the total number of dialogues in the test set. The indicator $1_{\text{correct\_intent}_i}$ is 1 for a match and 0 for a mismatch.
- **Parameter Accuracy.** This metric's evaluation by fifty users focuses on three criteria: key term inclusion, condition adherence, and negative vocabulary check, weighted as 4:3:3.
- **Structural Conformity.** Similarly, this metric is evaluated by human users, focusing on three primary aspects: whether the output conforms to the specified JSON format and the absence of repetition or garbled text, with weights allocated as 6:4.

Component Analysis. We conduct an ablation study on the component, as detailed in Table 2. Our findings indicate: (i) Compared to Qwen-VL-zero-shot, the model's performance after training has greatly improved in all metrics, with over 10% increase, especially in Structural Conformity, highlighting the fine-tuned model's strength in recognizing intent and parameters. (ii) The standalone SFT strategy occasionally generates outputs with uninterpretable repetitions, garbled text, or negative sentiments. Models trained with DPO regard such responses as suboptimal and suppress them *(Line 3 vs. Line 4)*. (iii) For DPO strategy, $IRPEM_{(SFT)}$ is a better base model than Qwen-VL-zero-shot. As it minimizes the gap between model outputs and preference data, leading to responses more attuned to human preferences *(Line 2 vs. Line 4)*.

Effectiveness Analysis. In the context of multi-turn dialogue for image generation and editing, we analyze these strategies, with the results presented in Table 3. This highlights the significance of these strategies in enhancing image editing during multi-turn conversations. Notably, the $IRPEM_{(RLHF)}$ integrated with Pipelines outperforms standalone DPO and SFT strategies by 2.12% and 4.30% in the IEC metric, showing significant gains over zero-shot strategies. This indicates our method has higher-quality images in the context of multi-turn dialogues.

5 Conclusion

In this paper, we introduce DialogDraw, a conversational image generation and editing system that also supports VQA and text dialogue functionality. Our approach stands out by accurately understanding and structuring users' natural language descriptions, enabling seamless integration with various mainstream open-source drawing models and pipelines. The system's high scalability is driven by continuous iterations of our proposed intent understanding model. We developed a multimodal dataset featuring multi-turn dialogues with rich drawing instructions, including image generation and editing. This dataset, curated from top pipelines in the open-source community, is used to train our intent recognition model. Experimental results show that DialogDraw outperforms current mainstream dialogue-based drawing models in intent recognition accuracy, drawing continuity, and consistency, as evidenced by both qualitative and quantitative results. However, our system has some limitations. Currently, it integrates only the 20 most popular pipelines and models, and we have made limited progress in enhancing VQA and text dialogue capabilities. In the future, we plan to expand the system's drawing capabilities and improve intent accuracy through human feedback to better align with user preferences.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; Manassra, W.; Dhariwal, P.; Chu, C.; and Jiao, Y. 2023. Improving Image Generation with Better Captions.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Shen, Y.; Gao, J.; Liu, J.; and Liu, X. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8721–8729.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Cheng, Y.; Gan, Z.; Li, Y.; Liu, J.; and Gao, J. 2020. Sequential attention GAN for interactive image editing. In *Proceedings of the 28th ACM international conference on multimedia*, 4383–4391.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Civitai. 2022. Civitai. https://civitai.com/.

Cui, X.; Li, Z.; Li, P.; Hu, Y.; Shi, H.; and He, Z. 2023. Chatedit: Towards multi-turn interactive facial image editing via dialogue. *arXiv preprint arXiv:2303.11108*.

Ge, Y.; Zhao, S.; Zeng, Z.; Ge, Y.; Li, C.; Wang, X.; and Shan, Y. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Li, H.; Hu, H.; et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12709–12720.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, M.; Long, Y.; Deng, X.; Chu, R.; Xiong, J.; Liang, X.; Cheng, H.; Lu, Q.; and Liu, W. 2024. DialogGen: Multi-modal Interactive Dialogue System for Multi-turn Text-to-Image Generation. *arXiv preprint arXiv:2403.08857*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*.

OpenAI. 2022. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenArt. 2022. OpenArt. https://openart.ai/.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E. M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wu, S.; Jiang, X.; and Wang, Y. 2023. DialogPaint: A Dialog-based Image Editing Model. *arXiv preprint arXiv:2303.10073*.

Yu, Y.; Zuo, S.; Jiang, H.; Ren, W.; Zhao, T.; and Zhang, C. 2020. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. *arXiv preprint arXiv:2010.07835*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.