DialogPaint: A Dialog-based Image Editing Model

Jingxuan Wei[2,3†] , Shiyu Wu[2,4†], Xin Jiang[1] , Yequan Wang[1*]
[1]Beijing Academy of Artificial Intelligence, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4]Institute of Automation, Chinese Academy of Sciences, Beijing, China
weijingxuan20@mails.ucas.edu.cn, wushiyu2022@ia.ac.cn
jiangxin@baai.ac.cn, tshwangyequan@gmail.com

Abstract

We present DialogPaint, an innovative framework that employs an interactive conversational approach for image editing. The framework comprises a pretrained dialogue model (Blenderbot) and a diffusion model (Stable Diffusion). The dialogue model engages in conversation with users to understand their requirements and generates concise instructions based on the dialogue. Subsequently, the Stable Diffusion model employs these instructions, along with the input image, to produce the desired output. Due to the difficulty of acquiring fine-tuning data for such models, we leverage multiple large-scale models to generate simulated dialogues and corresponding image pairs. After fine-tuning our framework with the synthesized data, we evaluate its performance in real application scenes. The results demonstrate that DialogPaint excels in both objective and subjective evaluation metrics effectively handling ambiguous instructions and performing tasks such as object replacement, style transfer, color modification. Moreover, our framework supports multi-round editing, allowing for the completion of complicated editing tasks.

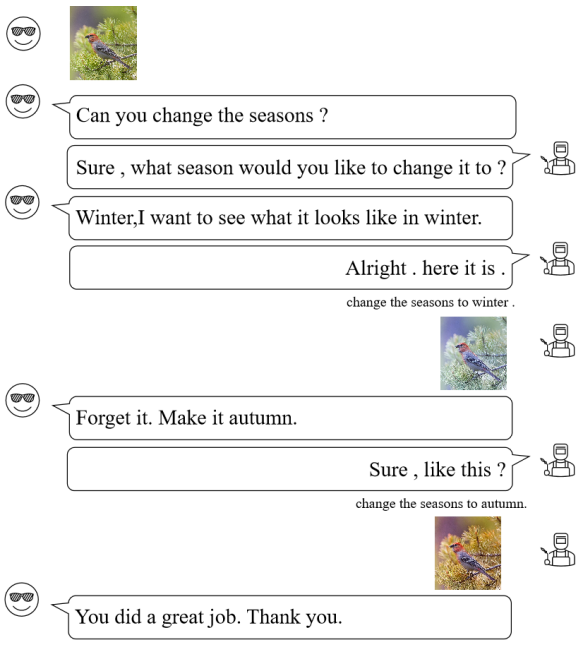


Figure 1: An example of interactive editing

1 Introduction

Recently, great progress has been achieved in the field of image generation through the use of diffusion models (Ho et al., 2020; Nichol and Dhariwal, 2022; Song et al., 2021; Dhariwal and Nichol, 2021; Rombach et al., 2022). These large-scale text-to-image models have enabled the synthesis of high-quality, diverse images using concise textual prompts. Owing to their vivid output and stable training performance, diffusion models have surpassed generative adversarial networks (GAN)(Goodfellow et al., 2020) in popularity. Consequently, an increasing number of individuals are utilizing various diffusion models for a wide array of tasks, engaging in the creation of personalized images.

Although image generation has become increasingly accessible to the general public, other tasks, such as image editing, remain considerably challenging.

This preference can be explained by the continuity of thought, as altering a portion of a scene aligns more closely with human cognition. Utilizing natural language instructions to edit images has proven to be both helpful and appealing to users. Current methods for semantic image editing using generative models are abundant and diverse (Zhan et al., 2021). However, these methods often prove unsuitable for human-computer interaction or rely heavily on personal adjustments. Based on an in-depth analysis of this issue, we identified two primary limitations that significantly impact the editing performance of these models.

*Corresponding author
†Equal contribution.

First, most text-to-image models struggle to process human instructions, a phenomenon we term "instruction unfriendly." This arises because these models are predominantly trained on declarative sentences, making imperative sentences, such as instructions, unfamiliar to them. Second, individuals frequently provide models with ambiguous instructions, which can lead to confusion regarding sources and targets in the absence of a given image. Some users simply employ vague phrases like "something else" to refer to the target, which existing models find difficult to interpret. To address these challenges, we propose using dialogues to clarify instructions. By engaging in conversation with large language models, precise commands can be extracted and subsequently employed to guide image generation.

In this study, we introduce a user-friendly approach to image editing through natural language conversation. Upon providing our model with an input image, users can engage in a dialogue with the model to convey their editing requirements. Our model is capable of discerning user needs and formulating a concise instruction to edit the image accordingly. We employ two models, a dialogue model (Shuster et al., 2022) and a image generation model (Rombach et al., 2022) to complete this task. The dialogue model is used to converse with the user and, if the user's input instruction is deemed ambiguous, it seeks clarification through additional questions. Finally, the language model generates a clear instruction, which the image generation model utilizes, along with the input image, to produce a new, edited image. Since there are no suitable existing datasets for fine-tuning these two large models, we adopted the approach of self-instruct (Wang et al., 2022) and generate simulated dialogues and image pairs for fine-tuning purposes. Results demonstrate that our model achieves zero-shot generalization in real-world application scenarios. Our model is capable of performing various edits, including object replacement, style transfer, and color alteration. A demo of our project is made publically available, with an interactive editing example presented in Figure 1.

2 Related Work

2.1 Large Language Models

Large language models (Radford et al., 2018; Shuster et al., 2022; Ouyang et al., 2022; Brown et al., 2020) is able to chat with humans fluently, which have been widely studied in recent years. These models, such as GPT-3 (Brown et al., 2020), have the ability to generate simulated data according to given samples, which is a convenient way to gather language data in a specific format and fine-tune other language models. Furthermore, conversation-oriented language models have also received attention.

DialoGPT (Zhang et al., 2020) is an open-domain conversation model that generates high-quality human-like responses in conversations. It uses large-scale generation models to achieve this. Meena (Adiwardana et al., 2020) is a chatbot developed by Google that aims to train more conversations and empathy in human interactions by utilizing large-scale conversation models. BlenderBot (Roller et al., 2021) is a conversation agent trained on different conversation datasets from social media platforms, and it can converse on a wide range of topics. ChatGPT (`https://openai.com/blog/chatgpt`) is a large-scale language model developed by OpenAI for generating high-quality text in a conversational context. It has shown exceptional performance on various conversational tasks.

Currently, BlenderBot and its subsequent open-source versions are popular in the field of conversation, while ChatGPT only has an API port available for now.

2.2 Diffusion Models

Diffusion model (Ho et al., 2020; Nichol and Dhariwal, 2022; Song et al., 2021; Dhariwal and Nichol, 2021; Rombach et al., 2022) is a new kind of generative models which generate images from Gaussian noise by progressively denoising it. The model gradually adds noise to the input image by a preseted noise adding method which is named as forward process. And then it uses a deep neural network to restore the original image, which is called sampling process. As the dimensions of latent space in diffusion models can be really high, the output images can be very fantastic with high quality and diversity. Usually, a diffusion model is trained on the following variant of the variational bound:

$$L_{simple} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{c},\boldsymbol{\epsilon},t}\left(||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_0, t, \boldsymbol{c})||_2^2\right) \quad (1)$$

where $\mathbf{x}_0$ and $\boldsymbol{c}$ are input images and optional conditions, $t \sim \mathcal{U}(0,1)$ are time steps and $\boldsymbol{\epsilon} \in \mathcal{N}(0,1)$ are added gaussian noise in forward process. $\boldsymbol{\epsilon}_\theta$ is

a learnable neural network which predicts the noise added on the image of previous moment. Usually a UNet (Ronneberger et al., 2015) is used to quickly and efficiently do this job. conditions $\boldsymbol{c}$ are the semantic embeddings of the input sentences processed by CLIP (Radford et al., 2021). With sampling methods such as DDIM (Song et al., 2021) and DPM-Solver (Lu et al., 2022a,b) that can speed up the sampling process Conspicuously, diffusion models is able to synthesize an image in $15 \sim 25$ steps. The widly used stable diffusion applies an AutoEncoder, which encodes the input image $\mathbf{x}$ into a latent space first and decode the sampling output to an real image. In this case, diffusion models will not deal with high-frequency details and can synthesis images with higher quality.

2.3 Text-driven Image Editing

Text-driven editing with GAN (Brock et al., 2019; Karras et al., 2021a, 2019; Abdal et al., 2020, 2021) has been carefully studied in recent years. Early works targeted at single task like style transferring. They trained the model with special image pairs to complete special editing tasks, which is based on domain transferring. With the advent of CLIP, now people can guide image editing with texts as input conditions. As for diffusion models, some of them (Ramesh et al., 2022; Saharia et al., 2022) natively have the ability for editing images due to the strong capabilities of text features extraction by CLIP(Radford et al., 2021). Models with mask guidance (Couairon et al., 2022) can make the edit more accurately. Another editing way is using textual inversion (Gal et al., 2022; Ruiz et al., 2022). Models will learn a speical word in textual embedding space and bind it with the specific subject in the given image. After training, with a sentence that contains the special word, the diffusion model can generate images of the specific subject in different scenes described by the sentence.

3 Methodology

We propose a new task: dialogue-based image editing. Our system receives images and editing instructions from users in the form of dialogue. The system actively solicits clarifications and provides feedback and summaries to complete the dialogue-based image editing. For this task, we first introduce a multi-turn dialogue-based image editing dataset (Sec. 3.1). Then, based on the generated dataset, we construct the dialogue-based image editing model (Sec. 3.2).

3.1 Construction of Dialogue and Image Editing Datasets

3.1.1 Building Dialogue Dataset

The construction of the dialogue and image editing datasets is shown in Figure 2, which includes Building Dialogue Dataset and Building Image Editing Dataset. For Building Dialogue Dataset, we randomly selected image captions from CUB-200-2011 (He and Peng, 2020), Microsoft COCO (Lin et al., 2014), DeepFashion (Liu et al., 2016), and Flickr-Faces-HQ (FFHQ) (Karras et al., 2021b) datasets, and combined them with prompt instructions to generate the necessary dialogue data. For Building Image Editing Dataset, we randomly selected image-text pairs from the same datasets, and input the texts into text-davinci-003 to generate editing instructions. These datasets were chosen due to their diversity and prevalence in the computer vision community. Regarding the construction of dialogue datasets, we randomly selected 10,000 image captions from four different datasets: CUB-200-2011, Microsoft COCO, DeepFashion, and FFHQ. Using self-instruct (Wang et al., 2022), we combined these image captions with prompt instructions and input them into text-davinci-003 to generate the necessary dialogue data. The overall process of constructing the dialogue dataset is shown in Step 1 of Figure 2.

As shown in the example in Figure 3, we first defined a Prompt Head to describe the dialogue generation task. Here, we instructed text-davinci-003 to "generate a dialog about a user ordering the system to edit data of the image based on the given image caption." Then, we randomly selected 20 manually written dialogue examples from a sample library containing 500 dialogue examples, as shown in the Example section of Figure 3. The Example section consists of two parts: the "Caption" which is the input image caption, and the "Dialog" which is the desired dialogue that includes simulating multiple rounds of conversation for modifying the image's colors, scenes, and other changes, as well as correcting fuzzy instructions. Finally, we concatenated an image caption after the Examples section and input the whole string into text-davinci-003. The response generated by text-davinci-003 was used to create the desired dialogue data.

Using the dialogue dataset construction method described above, we obtained a total of 10,000

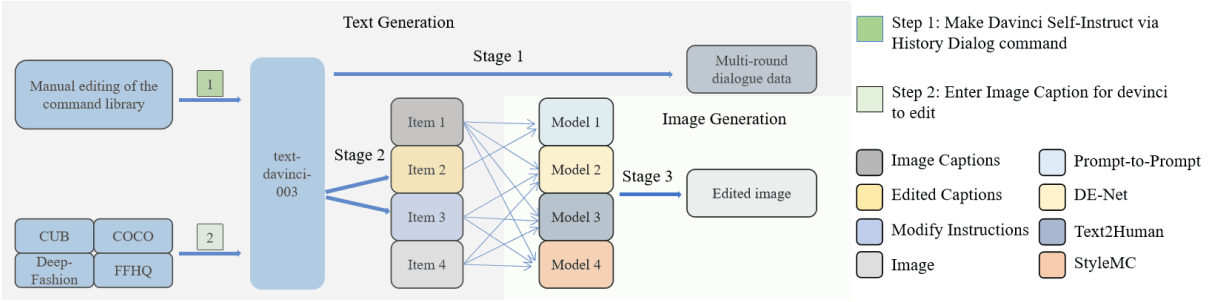


Figure 2: The processing of dialogue and image editing dataset construction

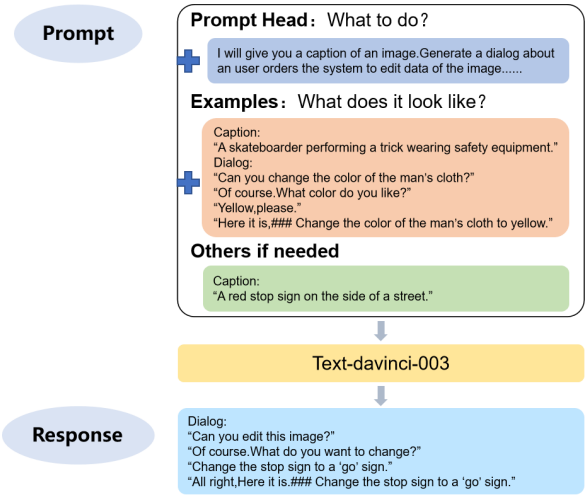


Figure 3: Example of Dialogue Dataset Construction

dialogue data samples that meet the needs of open-domain dialogue image editing, including various modifications to people, objects, backgrounds, etc. mentioned in the image captions.

3.1.2 Building Image Editing Dataset

The overall process for building the image editing dataset is shown in Step 2 and Stage 2 in Figure 2. The process is divided into two parts: Step 2 generates image editing instructions, and Stage 2 uses existing text-to-image editing models to generate edited images based on the text data generated in Step 2. In Step 2, we randomly selected 10,000 image-text pairs from the CUB-200-2011, Microsoft COCO, DeepFashion, and FFHQ datasets. Here, the text refers to the image caption, and we used self-instruct(Wang et al., 2022) to input the prompt instructions and image captions together into the text-davinci-003 model to generate the desired text editing instructions. The generated data was then fed into the text-to-image editing model in Stage 2 to obtain the edited images.

In Step 2, similar to Figure 3, we first defined a Prompt Head to describe the task of generating the dialogues. We told the text-davinci-003 model, "The following is the automatic generation of modification instructions based on caption, as well as the generation of new sentences to add modification instructions. Modification instructions are not limited to human editing, but can be any object, anything...". We then randomly selected 20 human-written editing instructions from a sample library containing 500 editing instruction examples. An example of such an instruction is shown in Figure 4 as "Example", which is divided into three parts: "Image Caption" as the input image caption, "Modify Instructions" as the editing instructions for modifying anything, and "Edited Captions" as the output generated by using the editing instructions as prompt to modify the caption. For example, given the image caption "A cat laying on top of a wooden bench.", the Modify Instructions "Change the cat to a yellow labrador.", the Edited Captions would be "A yellow Labrador laying on top of a wooden bench." Finally, we concatenated an image caption with the Examples, inputted it into the text-davinci-003 model, and obtained the desired Modify Instructions and Edited Captions data as the response.

Note that we generated some datasets here for partial transformation and object isolation of images to be used for fine-tuning the model in Section 3.2.2.

In Stage 2, we organize the editing instruction data generated in Step 2 and use multiple pre-trained models with Image Captions, Modify Instructions, Edited Captions to generate edited images. Inspired by previous works, we use four text-to-image editing models, including Prompt-to-Prompt (Hertz et al., 2022), DE-Net (Tao et al., 2022), Text2Human (Jiang et al., 2022), and StyleMC (Kocasari et al., 2022). As shown in Figure 2 Stage 2, Prompt-to-Prompt takes Image

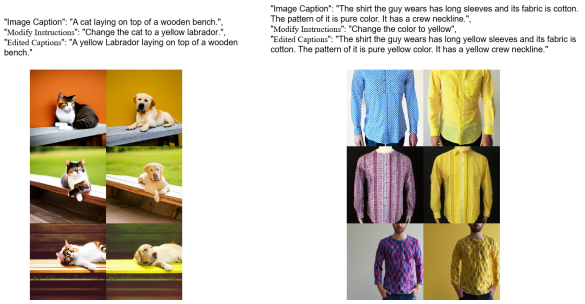


Figure 4: Example of Image Editing Dataset

Captions and Edited Captions as input to generate the original and edited images simultaneously, as it cannot directly modify the editing instruction. DE-Net, Text2Human, and StyleMC take Image Captions, Modify Instructions, and the original image as input to generate the edited image. After generating images using these four models, we filter them using the CLIP-based metric that measures the similarity change between the original and edited images. This approach maximizes the reduction of noise in the dataset and ensures data quality.

Finally, after filtering, we obtain a total of 6468 pairs of original image-text and edited image-text pairs, which satisfies the requirements of common open-domain image editing tasks.

3.2 Construction of Dialogue and Image Editing Models

The Construction of Dialogue and Image Editing Models module consists of two parts: Dialogue Model Construction and Image Editing Model Construction. We fine-tuned the BlenderBot and Stable Diffusion models, respectively, on our newly generated Dialogue Editing dataset. The BlenderBot model was used for generating context-aware dialogue responses, while the Stable Diffusion model was used for image editing tasks based on explicit textual instructions. The module aims to facilitate the generation of more realistic and coherent dialogues, as well as the generation of high-quality edited images based on natural language instructions. This module addresses the challenges of dialogue and image editing tasks and provides a practical solution for generating high-quality content.

3.2.1 Dialogue Model Construction

Our task is to generate open-domain conversations for image editing, where the model takes in a natural language prompt describing the image and an editing task, and generates a response dialogue that leads to the final edited image. This is achieved through a single or multiple rounds of dialogue, resulting in a clear instruction that guides the image editing module.

To accomplish this task, we fine-tuned the Blender dialogue model (Roller et al., 2021). The response generation is formulated as maximizing the probability of generating a response given the prompt, which can be expressed as:

$$P(\text{response}|\text{prompt}) = \prod_{i=1}^{n} P(w_i|w_{<i}, \text{prompt}) \tag{2}$$

where w_i denotes the i-th word in the response, $w_{<i}$ denotes the words before w_i, and n is the length of the response.

During training, we minimize the negative log-likelihood loss:

$$\mathcal{L} = -\log P(\text{response}|\text{prompt}) \tag{3}$$

To fine-tune our model for the task, we utilized our own dialogue dataset consisting of multi-turn dialogues for image editing. The objective of fine-tuning is to generate high-quality responses given a dialogue history x, with the aim of producing explicit editing instructions. Response generation is formulated as maximizing the probability of generating a response y given a dialogue history x, which can be expressed as:

$$\arg\max_y P(y|x) = \arg\max_y \prod_{t=1}^{T} P(y_t|y_{<t}, x) \tag{4}$$

where T is the length of the generated response, and $y_{<t}$ denotes the previously generated tokens.

3.2.2 Image Editing Model Construction

Our image editing model is fine-tuned based on the Stable Diffusion model (Hertz et al., 2022) to perform image editing according to explicit instructions from a dialogue model. We use the diffusion model to fine-tunes a pre-trained stable diffusion model on an image editing dataset to learn a network that predicts the noise to be added to the latent image based on text instructions. Specifically, we minimize the following diffusion target:

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \mathcal{E}\left(c_I\right), c_T\right)\|_2^2\right] \tag{5}$$

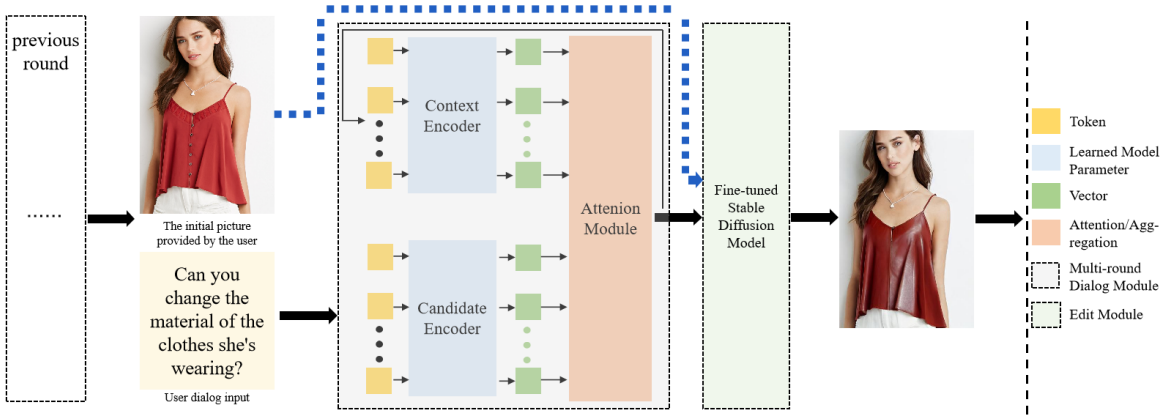


Figure 5: Model Architecture for Dialogue-Based Image Editing

Here, E and D are the encoder and decoder of a pre-trained variational autoencoder in the stable diffusion model, x is the input image, c_I is the image adjustment, and c_T is the text instruction adjustment. The diffusion process injects noise $z_t = E(x)$ into the encoded latent image, producing a noise latent image z_t with increasing noise levels over time steps $t \in T$. Given the conditions c_I and c_T, the network θ predicts the noise added to the noise latent potential z_t.

In addition, As with InstructPix2Pix, we added an extra input channel in the first convolutional layer, connecting z_t and $E(c_I)$ to support image adjustment. Unconditional diffusion guidance was also introduced, using two guidance scales s_I and s_T to adjust the trade-off between the correspondence with the input image and the editing instruction. The modified score estimation is given as follows.

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) =& e_\theta(z_t, \varnothing, \varnothing) \\ &+ s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ &+ s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned} \tag{6}$$

4 Experiments

4.1 Experimental Setup

We applied our model on two sets of data: 10,000 dialogue samples and 6,468 filtered image editing samples. For the dialogue model, we used 9,000 samples for training, 500 for validation, and 500 for testing. For the image editing model, we used 5,868 samples for training, 300 for validation, and 300 for testing. We fine-tuned BlenderBot using the version described in (Shuster et al., 2022). on the dialogue data, and fine-tuned our Stable Diffusion using the model provided by https://huggingface.co/timbrooks/instruct-pix2pix/tree/main on the image editing data. Both fine-tuning processes were performed on 8 Nvidia Tesla A100 40G GPUs.

When fine-tuning BlenderBot, we set the batch size to 128, embedding size to 2560, ffn size to 10240, n-heads to 32, n-positions to 128, n-encoder-layers to 2, n-decoder-layers to 24, dropout to 0.1, used the Adam optimizer with a learning rate of 7e-06, and implemented early stopping if there was no improvement for 10 consecutive iterations. For Stable Diffusion, we set the batch size to 32, input image size to 256, and kept the diffusion model configuration and no-classifier guidance algorithm the same as in the original paper. We fine-tuned the model for 125 epochs.

Finally, we fixed the parameters of the fine-tuned dialogue and image editing models and connected them according to the architecture shown in Figure 5, resulting in the implementation of our dialogue-based image editing system.

4.2 Qualitative Analysis of Experimental Cases

In order to evaluate the performance of our proposed dialog-based image editing model, we compared it to InstructPix2Pix, which is the baseline model in this paper, under the same precise single-turn instructions, and compared the results of the two models. The comparison results are shown in Figure 6. It can be observed that for InstructPix2Pix, when the input image undergoes significant transformations, the model exhibits overfitting problems, resulting in the loss of fine-grained background information from the original scene when changing the scene. For DialogPaint, the model can retain fine-grained background information while

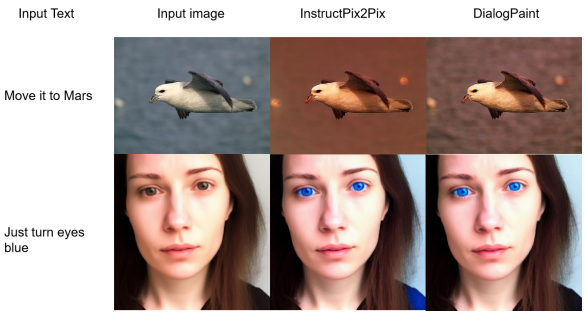


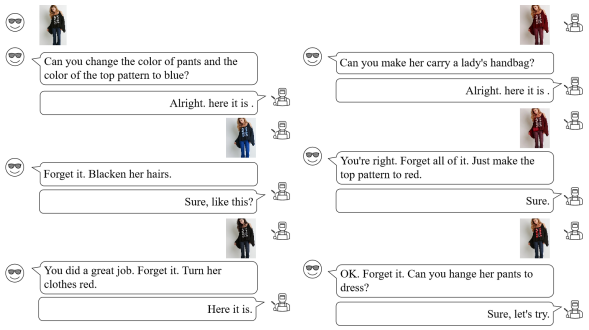
Figure 6: Comparison of Performance between DialogPaint and InstructPix2Pix under Same Precise Editing Instructions

completing the image transformation. In addition, for InstructPix2Pix, the model has difficulty isolating the specified object. For example, when given the instruction "Just turn eyes blue" the model cannot isolate the specified object, resulting in not only the eyes of the person but also a corner of their clothing becoming blue. When using DialogPaint, better object isolation can be achieved.

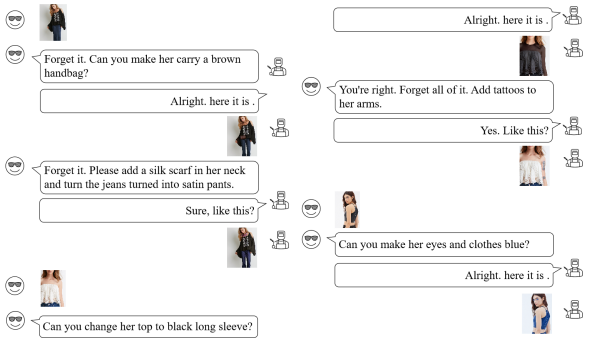
We attribute the model's ability to the facts that we avoid over-fitting and isolate specified objects in the dataset we provided. Our dataset contained more fine-grained images, and the descriptions were clearer, making it easier for the model to learn fine-grained knowledge and achieve precise transformation.

To further verify the ability of our proposed dialogue-based image editing model, we conducted multiple rounds of dialogue-based testing, as shown in Figure 7. We conducted over 15 rounds of image transformations, including continuous transformations and mid-image replacement according to the given instructions. In Figure 7a, we started with a fashion image and changed the color of different parts of the person's body and clothing, and added items to the image, achieving continuous editing. In Figure 7b, we continued the previous dialogue and made more fine-grained changes, such as changing the style and color of the person's clothing. In Figure 7c, we continued the previous dialogue and tried to delete and modify objects in the original image, fully demonstrating the precise editing capability of our model through dialogue.

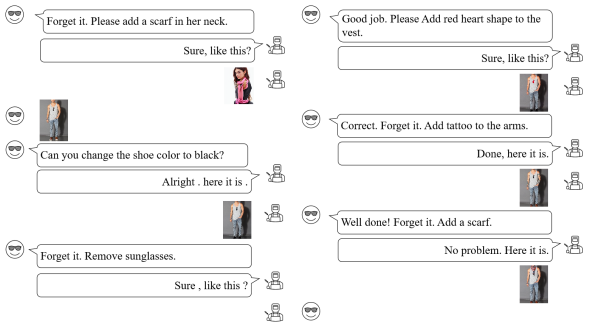
Furthermore, to further verify the dialogue-based editing capability of our model, we conducted multi-domain and multi-round testing in different fields, such as animals (Figure 8a), scenes (Figure 8b), and fruits (Figure 8c). The results showed



(a) Dialogue-based Fashion Transformation 1



(b) Dialogue-based Fashion Transformation 2



(c) Dialogue-based Fashion Transformation 3

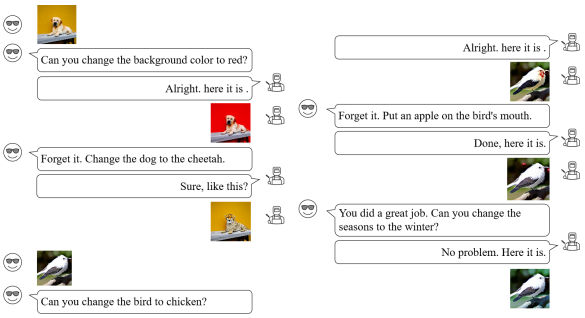
Figure 7: Fashion Transformation using Dialogue-guided Image Editing Model.In each round of the example shown, we demonstrate various image editing instructions using clear and specific commands.

that our model can achieve precise image editing through dialogue in different domains, meeting user expectations.

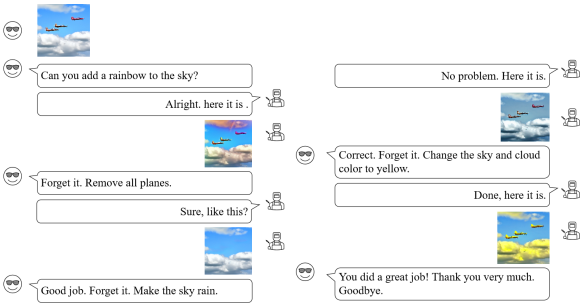
4.3 Quantitative Analysis of Evaluation Metrics

Based on the evaluation of our proposed dialogue-based image editing model, we have used a combination of objective and subjective metrics to assess its performance. The results are shown in Table 1.

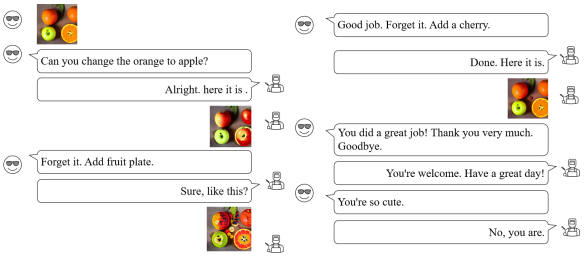
For objective metrics, we have used perplexity (ppl), Fréchet Inception Distance (FID), and Precision and Recall Distance (PRD). Perplexity measures the model's ability to predict the next word in



(a) Animal Transformation using Dialogue-based Image Editing Model



(b) Scene Transformation using Dialogue-based Image Editing Model



(c) Fruit Transformation using Dialogue-based Image Editing Model

Figure 8: Multi-modal Image Editing using Dialogue-based Approach. Here, various editing operations are demonstrated using explicit editing instructions.

a dialogue sequence, with lower values indicating better performance. FID compares the generated images to real images based on their statistics, with lower values indicating better image quality. PRD compares the distributions of generated and real images in a feature space, with lower values indicating a better match between the distributions.

For subjective metrics, we asked 100 participants to use the dialogue editing model and rate their overall satisfaction on a scale of 1 to 5. The average rating for overall satisfaction was 4.22, indicating a high level of satisfaction with the model's performance. Additionally, we asked participants to rate the quality of the generated images on a Mean Opinion Score (MOS) scale of 1 to 5, and the average MOS was 4.32, indicating a high level

Table 1: Evaluation Metrics for the Dialogue-based Image Editing Model

Evaluation Metric	Score
Perplexity (ppl)	1.578
Fréchet Inception Distance (FID)	1.52
Precision-Recall Distance (PRD)	1.56
Overall Satisfaction	4.22
Mean Opinion Score (MOS)	4.32

Figure 9: Qualitative and Quantitative Comparison of Performance Improvement through Image Editing before and after Fine-tuning

of image quality.

Overall, our dialogue-guided image editing model demonstrates its high performance in terms of both objective and subjective evaluation metrics, which indicates its potential for real-world applications.

4.4 Model Limitations

We introduced the task of dialog-based image editing and demonstrated the capability of our proposed dialog-based image editing model. Although our method can achieve image editing through dialog by providing explicit instructions, removing ambiguous instructions, summarizing context, and making stylistic, color, and other changes to the image, there are still some limitations.

Due to the limited number of dialog samples and image editing operations in the current dataset, our model exhibits some limitations when dealing with complex dialog-based image editing tasks. For example, when explicit instructions are given through dialog, the image editing effect may be unsatisfactory due to the complexity of the original image content, as shown in the failed examples in Figure 9.

In Figure 9, we present some examples of insufficient precision in fine-grained changes, such as

in the top left corner, where the instruction is to change the color of the upper body's tank top to black, but the color of the entire body is changed instead. This indicates that local color processing is still not sophisticated enough.

We will continue to improve the precision of the dialogue-based image editing instructions and the image editing module to gradually overcome these limitations.

5 Conclusion

In this paper, we present a dialogue-based image editing model that enables image modification through explicit instructions in a conversation. To facilitate this, we constructed a dataset containing both dialogue and image editing, and conducted fine-tuning using dialogue and image generation models. Our experimental results indicate that the proposed model exhibits strong performance in both objective and subjective evaluation metrics, showcasing its dialogue-based image editing capabilities across various domains. However, due to the limited number of dialog samples and image editing operations present in the dataset, the model currently faces challenges in handling complex editing tasks. Moving forward, we plan to enhance the performance of both the instruction and image editing components in order to progressively mitigate these limitations.

In future research, we will further explore the potential of dialogue-based image editing models and attempt to apply them to a wider range of fields, such as smart homes and facial recognition. We will also endeavor to refine our dataset and collect more dialog samples with more diverse operations to enhance our model's performance. Furthermore, we will investigate other ways to combine our model with other models or technologies to achieve more efficient image editing and processing.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: How to edit the embedded images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8293–8302. Computer Vision Foundation / IEEE.

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3):21:1–21:21.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *CoRR*, abs/2211.09800.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *CoRR*, abs/2210.11427.

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *CoRR*, abs/2208.01618.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Xiangteng He and Yuxin Peng. 2020. Fine-grained visual-textual representation learning. *IEEE Trans. Circuits Syst. Video Technol.*, 30(2):520–531.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2human: text-driven controllable human image generation. *ACM Trans. Graph.*, 41(4):162:1–162:11.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021a. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2021b. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-based real image editing with diffusion models. *CoRR*, abs/2210.09276.

Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. 2022. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3441–3450. IEEE.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104. IEEE Computer Society.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022a. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *CoRR*, abs/2206.00927.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *CoRR*, abs/2211.01095.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alexander Quinn Nichol and Prafulla Dhariwal. 2022. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *CoRR*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021,*

Online, April 19 - 23, 2021, pages 300–325. Association for Computational Linguistics.

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*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CoRR*, abs/2208.12242.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, and Joshua Lane. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ming Tao, Bing-Kun Bao, Hao Tang, Fei Wu, Longhui Wei, and Qi Tian. 2022. De-net: Dynamic text-guided image editing adversarial networks. *CoRR*, abs/2206.01160.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *CoRR*, abs/2212.10560.

Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. 2021. Multimodal image synthesis and editing: A survey. *CoRR*, abs/2112.13592.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.