DialogGen: Multi-modal Interactive Dialogue System with Multi-turn Text-Image Generation

Minbin Huang[1§*], **Yanxin Long**[2*], **Xinchi Deng**[2], **Ruihang Chu**[1], **Jiangfeng Xiong**[2], **Xiaodan Liang**[3], **Hong Cheng**[1§], **Qinglin Lu**[2], **Wei Liu**[2]

[1]The Chinese University of Hong Kong [2]Tencent [3]Shenzhen Campus of Sun Yat-sen University

Abstract

Text-to-image (T2I) generation models have significantly advanced in recent years. However, effective interaction with these models is challenging for average users due to the need for specialized prompt engineering knowledge and the inability to perform multi-turn interaction, hindering a dynamic and iterative creation process. Recent attempts have tried to equip Multi-modal Large Language Models (MLLMs) with T2I models to bring the user's natural language instructions into reality. Hence, the output modality of MLLMs is extended, and the multi-turn generation quality of T2I models is enhanced thanks to the strong multi-modal comprehension ability of MLLMs. However, many of these works face challenges in identifying correct output modalities and generating coherent images accordingly as the number of output modalities increases and the conversations go deeper. Therefore, we propose **DialogGen**, an effective pipeline to align off-the-shelf MLLMs and T2I models to build a Multi-modal Interactive Dialogue System (MIDS) with the ability to perform multi-turn Text-to-Image generation. It is composed of drawing prompt alignment, careful training data curation, and error correction. Moreover, as the field of MIDS flourishes, comprehensive benchmarks are urgently needed to evaluate MIDS fairly in terms of output modality correctness and multi-modal output coherence. To address this issue, we introduce the Multi-modal Dialogue Benchmark (**DialogBen**), a comprehensive bilingual benchmark designed to assess the ability of MIDS to generate accurate and coherent multi-modal content that supports various users' need. It contains two evaluation metrics to measure the model's ability to switch modalities and the coherence of the output images. Our extensive experiments on DialogBen and user study demonstrate the effectiveness of DialogGen in producing correct output modalities and coherent multi-modal outputs compared with other State-of-the-Art models. DialogBen will be fully open-sourced and we hope it can contribute to the community for building more powerful MIDS.

1 Introduction

In recent years, Text-to-Image (T2I) diffusion models(Ho et al., 2020; Rombach et al., 2022; Song et al., 2020; Peebles and Xie, 2023) emerges as a focal point of interest. These models, which generate visual content based on textual input, have the potential to revolutionize various industries, including advertising, entertainment, and education (Zhang et al., 2017; El-Nouby et al., 2019). However, existing T2I models are still limited in comprehending user's complex natural language instructions due to the limited capacity of text encoders. This leads to a big challenge for users when interacting with T2I models. Besides, prompt engineering expertise is required to accurately bring the user's instruction into reality, which hinders its broader application to average users. Moreover, users may potentially generate images based on previous conversations and generation results, which calls for a multi-turn T2I generation. In response to this need, motivated by the strong language comprehension ability of Large Language Models (LLMs), an increasing number of researchers are turning their attention towards the integration of LLMs as a means to facilitate more natural and intuitive communication between users and T2I models. These works include Text-to-Image generation models equipped with open-source LLMs (Wu et al., 2023; Ge et al., 2023; Koh et al., 2023a; Zeqiang et al., 2023; Sun et al., 2023b; Koh et al., 2023b; Chen et al., 2023b) and closed-source LLMs such as DALLE3 (Betker et al., 2023) and RPG(Yang et al., 2024). This gives rise to a prosperous field of Multi-modal In-

*Equally contributed to this work.

§Affiliated with Department of Systems Engineering and Engineering Management, and Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Hong Kong

Findings of the Association for Computational Linguistics: NAACL 2025, pages 411–426

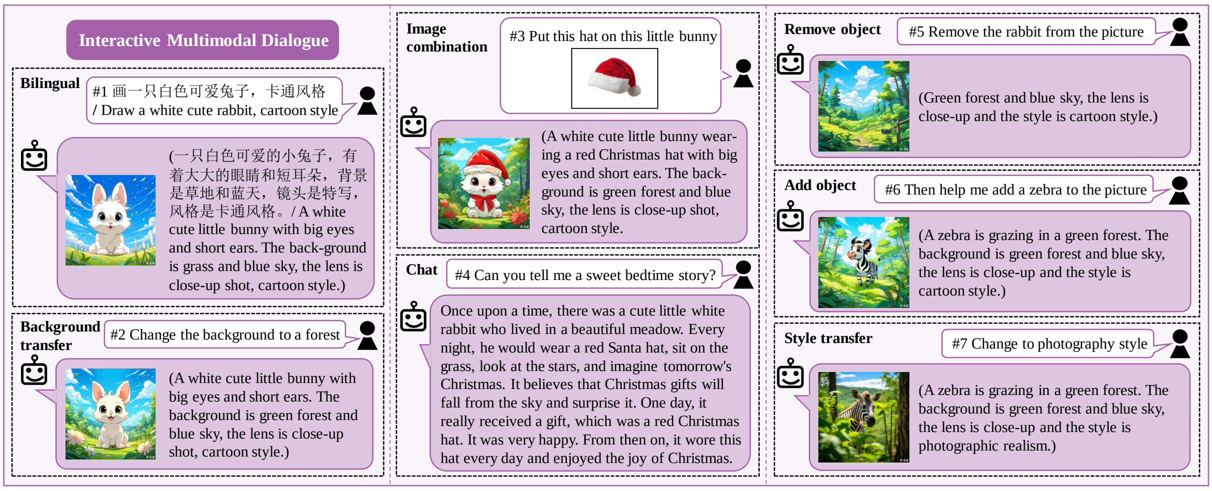


Figure 1: Our proposed DialogGen that can perform bilingual multi-turn multi-modal tasks responding to user's natural language instructions to meet various users' needs.

teractive Dialogue System (MIDS) that receives multiple modalities and outputs multiple modalities, as can be seen at the top of Fig. 3. However, combining MLLMs and T2I models to build a powerful MIDS is non-trivial. Experimental investigations reveal that contemporary solutions(Wu et al., 2023; Ge et al., 2023) face considerable challenges in accurately identifying users' intentions and bringing the users' instructions into reality. This is particularly different from the success of those MLLMs that process inputs from diverse modalities and solely generate outputs in text format (Chen et al., 2023b; Zhu et al., 2023; Liu et al., 2023b; Zhang et al., 2023; Dai et al., 2023). If the output encompasses multiple modalities, it presents a significantly higher degree of complexity than single-modality output. Therefore, to build a more powerful MIDS that connects MLLMs and T2I models, we propose DialogGen, an effective pipeline to align off-the-shelf MLLMs and T2I models for building MIDS. We first conduct drawing prompt alignment that transforms the output of the MLLM to the form that T2I models favor. Later, we finetune the MLLM with carefully curated bilingual modality switching data. Besides, motivated by the learning process of a student who iterative learns and corrects his own mistakes guided by the teacher, we imitate the error correction process to further boost the performance of our MLLM, where the generation results are more aligned with human intention. Our DialogGen is compatible with any current State-Of-The-Art (SOTA) T2I models.

As the field of MIDS continues to flourish, there is an urgent demand for a fair and comprehensive evaluation of such emerging systems. In response to this need, we introduce the Multi-modal dialogue Benchmark (DialogBen), an extensive benchmark to evaluate the capabilities of MIDS, which can engage in multi-turn dialogues with both input and output spanning multiple modalities. This benchmark contains the assessment of Modality Switching and Generation Coherence ability. It contains 9957 three-turn conversations covering 7 image editing types and 13 topic types, which is shown in Fig. 2. This benchmark aims to provide a proper assessment framework for the growing field of MIDS. By leveraging DialogBen, researchers can gain a deeper understanding of the challenges and opportunities in developing a more powerful MIDS that can switch between different output modalities, generate semantically coherent images, and ultimately lead to more versatile and user-friendly MIDS. Fig 1 illustrates the capabilities of MIDS built with our DialogGen. Comprehensive experiments on DialogBen have shown our superiority of DialogGen over current SOTA models.

To conclude, our contributions can be summarized as follows:

- We propose DialogGen, an effective pipeline to build a multi-modal interactive dialogue system with multi-turn Text-Image generation ability. It is built upon drawing prompt alignment, careful bilingual training data curation including object consistency guarantee, and instruction tuning data mixing. Besides, we also propose to train on error correction data from stronger LLMs to learn from mistakes it has made.

- We propose DialogBen, a comprehensive multi-turn bilingual benchmark that contains 9957 multi-turn multi-modal conversations covering 7 editing instruction types and 13 topic types to evaluate MIDS in terms of modality switching ability , and the generation coherence.
- Comprehensive experiments show that MIDS built by DialogGen beats current SOTA models in terms of modality switching accuracy, generation coherence VQA score, and user preference.

2 Related Works

Prompting for Large Language Models. One of the most prominent strengths of Large Language Models (LLMs) is their in-context-learning(Dong et al., 2022; Sun et al., 2023a) ability. This enables users to effortlessly tailor LLMs for specific tasks or augment their capabilities using straightforward prompts. For instance, the chain-of-thoughts(Wei et al., 2022) technique pioneered the enhancement of LLMs by requesting them to produce a sequence of intermediate reasoning stages. Subsequently, various refined prompting methods have been developed, incorporating heuristics such as majority voting, backtracking, and graph-based thought representations. In our work, we rely on prompting GPT-4 for creating the benchmark dataset.

Multi-modal Large Language Models and Evaluations. Recent success of Large Language Models (LLMs) (Touvron et al., 2023a,b; Chowdhery et al., 2022; Zhang et al., 2022) has largely benefited research on natural language processing. Inspired by the strong generation ability, researchers try to improve visual comprehension and Text-to-Image generation ability utilizing the generalization power of LLMs, leading to the prosperity of research on Multi-modal Large Language Models (MLLMs). All MLLMs can receive inputs from multiple modalities; however, quite a few of them can produce the output from multiple modalities (Wu et al., 2023; Ge et al., 2023; Chen et al., 2023b; Zeqiang et al., 2023; Koh et al., 2023a), while others can only produce output in the text format. (Zhu et al., 2023; Liu et al., 2023b,a). MLLMs that support image generation generally fall into two categories: one with the intermediate prompt for feeding into the image diffusion model, the other directly outputs the embedding vector for the diffusion model and generates the required images. Our DialogGen falls into the first category like DALLE3, as we believe that a user-friendly intermediate result is advantageous for analyzing the model behavior and providing feedback for model improvement.

The prosperous multi-modality-out MLLMs call for an effective benchmark to evaluate their capability. Our DialogBen specializes in the setting that users repeatedly seek image generation outputs beyond casual conversations. While previous literature such as MMDialog (Feng et al., 2022) provides a large-scale dataset for multi-turn multi-modal conversations, their focus is more on generating text-based scenarios. Moreover, they lack evaluations of the multi-turn multi-modal output, because previous models often struggle to comprehend long conversations without the support of LLMs and MLLMs.

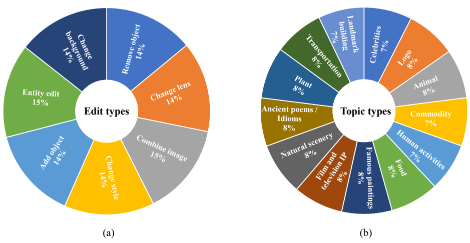


Figure 2: Our benchmark encompasses 7 edit instruction types and 13 topic types.

3 DialogBen

3.1 Benchmark Design

The benchmark is designed to consist of a diverse set of dialogue scenarios that involve Text-to-Image generation. It encompasses various topics and image editing types, regarding different modality-switching scenarios for each turn of conversation. To ensure broad coverage of conversational scenarios, we explore different combinations of input and output types based on four primary categories, i.e., $T \rightarrow T$, $T \rightarrow I$, I+$T \rightarrow T$ and I+$T \rightarrow I$, where I, T represent image and text respectively. We also summarize the data proportion (see Table.7 in appendix) for each modality-switching scenario in each round and find that they are basically balanced. By selecting a type for each conversation turn, we devise $4^3 = 64$ three-turn conversation compositions. Subsequently, we employ GPT4 to generate 'meta prompts' for these 64 types, each representing a unique conversational flow. For each conversation combination, we traverse 13 predefined topics and 7 image editing methods (see Fig. 2) to

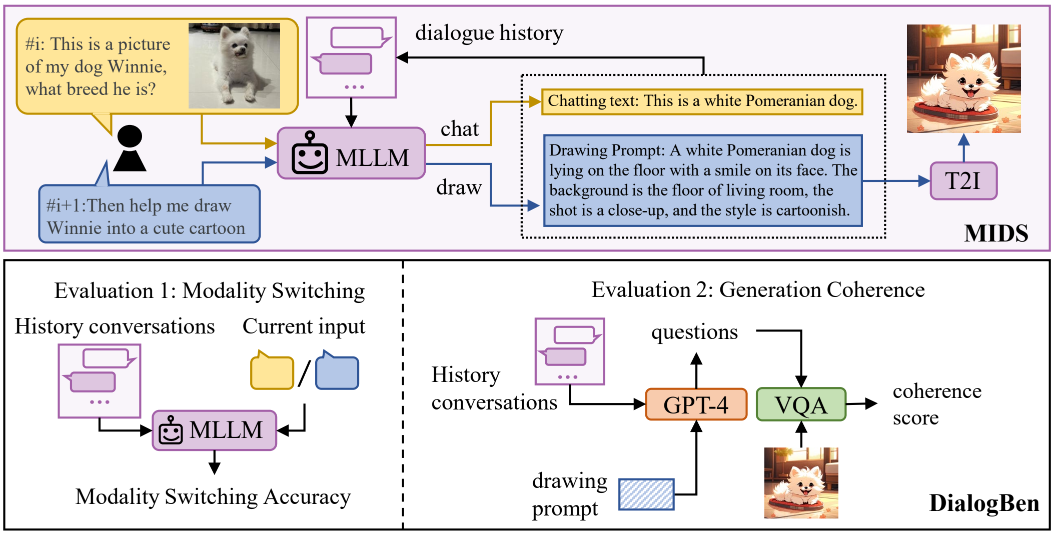


Figure 3: Overview of MIDS and DialogBen. MIDS can respond to the multi-modal user instructions with either a text response or a drawing prompt to be sent to a T2I model for image generation. DialogBen consists of 9957 three-turn multi-modal dialogs and two evaluation metrics to assess the capability of MIDS.

yield approximately 15k samples after query GPT4 using corresponding 'meta prompts'. To ensure the quality of the evaluation, we utilize the LLM model to filter out any mismatched modality output intentions between the input instructions and the ground truth label. Later, we split an evaluation set that for each combination there is at least one sample and hence get an evaluation dataset comprising 9957 three-turn multi-modal conversations, which are characterized by incorporating both images and text as the input and output. While the remaining $\sim$ 4.5k samples are used as the training set. We refer to the training set and evaluation set as DialogBen-train and DialogBen-test.

Example 1: Coherence VQA

User:
What kind of merchandise is this and what is it used for?
Assistant: This product is an oven, mainly used for baking or roasting food, with temperature control and timing functions, is a common appliance in modern home kitchens.
User: What if you change the color of the oven so that it fits more in with the kitchen decor, say silver?
Questions:
1. Is there an oven in the picture?
2. Is the oven placed in the middle of a kitchen?
3. Is the background of the picture a modern kitchen?
4. Is the shot of this picture panoramic?
5. Is the color of the oven in the picture silver?

3.2 Evaluation Metrics

3.2.1 Modality Switching Accuracy

Precisely identifying users' intentions and producing outputs in the suitable modality, whether with text or images, is essential for Multi-modal Interactive Dialog Systems (MIDS). In this way, the systems can deliver more relevant, informative, and contextually fitting responses. Therefore, we think it's important to check how well these systems can switch between modes of communication in our benchmark test. We propose to evaluate the Modality Switching accuracy for each turn and focus on image and text modalities, as shown at the left bottom of Fig. 3. When a user inputs an image, it is usually accompanied by text describing the user's query related to the image. The system's output typically consists of image and text modalities. For each round, we assess 4 different modality switching scenarios mentioned in Sec. 3.1, to ensure a comprehensive evaluation of the system's ability to generate the right multi-modal outputs. After inference on the benchmark, we calculate the modality switching accuracy as follows: suppose in the ith round, there are n_{ij} samples belonging to j^{th} modality switching scenario, let $c_{ij}^{(k)}$ be the binary variable, such that $c_{ij}^{(k)} = 1$ indicates the model producing the correct output modalities. Then, the Modality Switching Accuracy Acc_{ij} in the ith round for j^{th} scenarios is calculated by

$$Acc_{ij} = \left(\sum_{k=1}^{n_{ij}} c_{ij}^{(k)} \right) / n_{ij}$$

3.2.2 Generation Coherence

To evaluate the generation quality in multi-turn conversations in image generation mode, and to better compare with MIDS that produce images directly without intermediate text prompts, we propose a Generation Coherence VQA score.

Inspired by (Huang et al., 2023), we employ the visual question answering (VQA) capability of BLIP(Li et al., 2023) to assess the generation quality. As shown at the right bottom of Fig. 3, for each query in our DialogBen-test, we utilize the historical conversation context, prompting GPT-4 to generate the required elements for the desired output images. These elements are then transformed into a set of alternative questions and their corresponding answers. This approach equips DialogBen-test with the affiliated collections of simple questions about the desirable output images. In Example 1 we showcase a sample from the candidate sets of the questions to be asked.

During the evaluation, we employ the BLIP-VQA model (Huang et al., 2023) to process the images generated by candidate models and the alternative question. The probability of the correct answer to each question is used as the score. We then calculate the average score of all questions associated with the generated image to determine the image's score. The Generation Coherence VQA Score for the entire test set is found by averaging these scores across all images. Specifically, for each test query q_i affiliated with m_i questions and the generated image I_i, the generation coherence score S_i is formulated as $S_i = Mean(\{P(BLIP(q_i^{(j)}, I_i) = GT(q_i^{(j)}, I_i))\}_{j=1}^{m_i})$

4 DialogGen

In this section, we introduce the components of DialogGen, which include Drawing Prompt Alignment, Training Data Curation, and Error Correction. Given an off-the-shelf MLLM $M_C(I, T)$ and a T2I model G, DialogGen integrates them seamlessly to build a powerful MIDS. The overall framework can be seen in Fig. 4. We will introduce them part by part.

4.1 Drawing Prompt Alignment

Suppose the original training data of G is $D_G = (I_i, T_i) \sim P(I, T)$, where I and T represent image and text, respectively. The user's text instructions should be aligned with the distribution of D_G for optimal image generation. Yet, because D_G typically comes from various data resources, it is hard to get the ground truth of P.

However, since re-captioning has been shown effective in various works(Betker et al., 2023; Yang et al., 2024) for improving better T2I generation quality, we utilize a re-caption model for text refinement. For a given MLLM M_C, we first apply it to re-caption the training set of G to obtain a new training set $D'_G = (I_i, T'_i) \sim P'(I, T)$, where $T'_i = M_C(I_i, P_{cap})$ and P_{cap} is the caption prompt such as "Please describe this image in detail.". Then we re-train the T2I model on D'_G. Afterward, when querying GPT-4 for training data generation, we sample $(I'_{icl}, T'_{icl}) \sim P'(I, T)$ as the in-context learning samples in the "meta-prompts", aligning the generated prompts with P'. Thus, the training data on image generation would have a similar drawing prompt distribution as P', which is exactly what G is re-trained on. In this way, the transformed drawing prompts match the given T2I models and thus lead to the expected image generation results. If we could directly access the ground truth of P without re-captioning through M_C, we would simply draw samples (I_{icl}, T_{icl}) from $P(I, T)$ for in-context learning to guide data generation process.

4.2 Training Data Curation

Object Consistency Guarantee. In multi-turn image generation, our goal is to ensure that objects generated across different conversational turns remain as consistent as possible. To achieve this, we add the following constraints for training data generation: for image generation that builds upon the images produced in previous rounds, the transformed drawing prompts should satisfy the user's current demand while being altered as little as possible from the drawing prompts used for previous images. Fig.1 #6 and #7 show the minor differences between the prompts in consecutive turns. Moreover, during the inference phase of a given conversation, we fix the random seed of the T2I model. This approach significantly increases the likelihood of maintaining object consistency throughout the dialogue.

Instruction Tuning Data Mixing. To maintain the multi-modal conversation ability, we've included a range of open-sourced uni/multi-modal conversation datasets, denoted as D_o. Besides, to address the lack of datasets that turn the user's natural language input into detailed description prompts that fall into P', the suitable distribution of D_G for

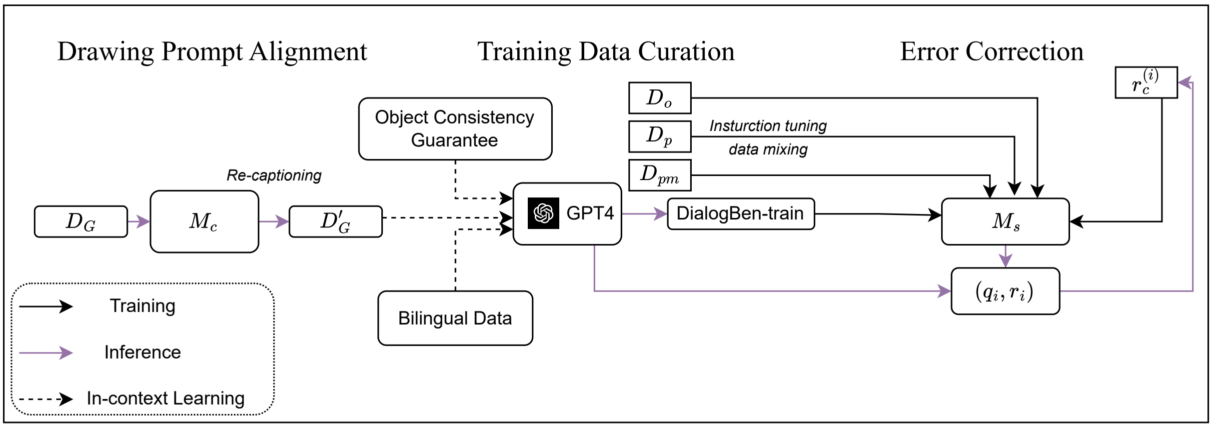


Figure 4: The overall pipeline of DialogGen which consists of Drawing Prompt Alignment, Training Data Curation, and Error Correction. The details of these components are elaborated in Sec.4

G to generate images, We first collect $n = 10$ manually-annotated labels. These labels will transform user ideas into semantically appropriate, detailed prompts that align with P'. Using these labels as in-context samples, we then query GPT-4 to generate a dataset that contains single-turn text-to-prompt samples, referred to as D_p. Since we aim at building a multi-turn dialogue system, we randomly shuffle and concatenate the single-turn samples from D_o and D_p to get a pseudo-multi-turn dataset D_{pm}. This dataset features multi-turn conversations not necessarily preserving semantic coherence, simulating the scenarios in which the user may switch the topic within a conversation. More importantly, we mix the collection of D_o, D_p, D_{pm} together with D_T, the training set of DialogBen. Please refer to Table. 5 in the appendix for detailed compositions of the instruction tuning data and training data samples.

Bilingual training Current MIDS typically only supports English scenarios and is trained on English data. Motivated by work (Chen et al., 2023a) that multilingual training can benefit LLMs' reasoning ability, we extend the training data from unilingual to bilingual. Specifically, we develop a Chinese counterpart of DialogGen, which is identical in structure to the original but differing in language. We refer them to DialogBen-en and DialogBen-cn, respectively. We conduct experiments on these versions individually as well as their mixed version. Please refer to Sec. 5 for more details.

4.3 Error Correction

Data Collection. To enhance our models' understanding of users' intentions, collecting more data is a common but often expensive approach. Instead, we propose learning from previous models' errors as a more efficient strategy. Specifically, given a training dataset D, we obtain a student model M_s after several training epochs. We then inference M_s on D to generate a collection of potentially noisy query-response pairs $(q_i, r_i), i = 1, 2, ..., |D|$. Next, we introduce a teacher corrector model M_T and a prompt P_f to query M_T, where P_f is the few-show prompt to instruct M_T to generate error correction data, as detailed in the Appendix. We employ the corrector model M_T to generate a correction c_i for each sample in the training set:

Table 1: Comparison of Modality Switching (MS) accuracy (%) of different models under standard fine-tuning and error correction counterparts with different style.

Model	Training	#Inference	DialogBen-MS	
			1st / 2nd / 3rd	Avg.
DialogGen-A	Standard Fine-Tuning	1-step	93.6 / 87.1 / 84.6	88.4
	+ Error Correction	1-step	96.9 / 95.6 / 93.4	95.3(+6.9)
	+ Error Correction	2-step	97.8 / 97.4 / 93.6	96.2(+7.8)
DialogBen	Standard Fine-Tuning	1-step	96.4 / 96.5 / 95.6	96.1
	+ Error Correction	1-step	97.1 / 97.9 / 96.7	97.2 (+1.1)
	+ Error Correction	2-step	97.3 / 97.9 / 96.9	97.4 (+1.3)

$$c_i \sim \mathcal{M}_T(P_f \oplus q_i \oplus r_i) \quad (1)$$

Note that in P_f we set up 3 rules for M_T to determine whether the output is satisfactory. Therefore, we include 3 human annotated error-correction examples as the few-shot in-context learning samples. In Example 4, there is an example showcasing the correction data generated by M_c.

Training. During training, we use a correction prompt P_c to prompt the student model to obtain the correction response $r_c^{(i)}$,

$$r_c^{(i)} \sim M_s(P_c \oplus q_i \oplus r_i) \quad (2)$$

The training loss is calculated between both the correction response and the provided suitable output. Note that we include both the feedback from correct and incorrect responses. Feedback from correct responses enhances the model's confidence in its accurate actions. On the other hand, feedback from incorrect responses alerts the model to its mistakes and how to avoid them in the future. By incorporating this error correction mechanism into MLLM, we would create a more robust and reliable system capable of generating high-quality, contextually appropriate responses. This approach not only helps the model to learn from its mistakes but also fosters a deeper understanding of the nuances and complexities involved in multi-modal interactive dialogue tasks.

5 Experiments

5.1 Models and Baselines

Base model. Note that we aim to build a strong bilingual Mutli-modal Interactive Dialogue System, so the base MLLM needs to have a considerable perception ability in both Chinese and English. Therefore, we carefully choose our base MLLM as Qwen-VL (Bai et al., 2023), which is an open-sourced MLLM that supports both Chinese and English input and output. It is capable of receiving images and texts as input and performing Multi-modal comprehension tasks like text reading, and image understanding in the format of text. For T2I models G, we conduct experiments on Hunyuan-DiT (Li et al., 2024), which is an open-source T2I model. This is because the re-captioning is performed on the training dataset of Hunyuan-DiT. For fair comparison in Coherence VQA score, we also conduct experiments for Stable Diffusion v1.5 and Stable Diffusion v2.1.

Baselines

Qwen-VL-few-shot : We provide a simplest way to equip Qwen-VL with T2I models G without training. We first add the description of the system we aim to build to serve as the system prompt during inference. Then we provide n manually designed samples demonstrating how to transform the multi-turn user instruction into drawing prompts to be sent to G, where $n = 0, 1$.

NExT-GPT(Wu et al., 2023): It is one of the earliest works to introduce the concept of any-to-any MLLMs. They introduce a projection layer to align the X-encoder (X can be image, audio, or video) and X-diffusion model with the LLM embedding space to build an MIDS that receives text, image, audio, and video both as candidate input and candidate output.

SEED-LLaMA(Ge et al., 2023): It designs a SEED tokenizer to tokenize the image and train the multi-modal discrete sequence in a unified objective. Therefore, when it needs to perform image generation, image tokens can be predicted autoregressively during inference and hence be decoded into images.

5.2 Training and Inference

To make it compatible with our goal of performing interactive multi-modal image generation, we add a modality switch token `<draw>` before the model output to identify the recognized intention of the user and a training prompt P_T as "Please first identify the intention of the user, if it is to draw please append `<draw>` before the output". To emphasize the effectiveness of D_T, we named the refer DialogGen-A as the model trained with $D_A = D_o \cup D_p \cup D_{pm}$, while DialogGen-B is the model trained with $D_B = D_o \cup D_p \cup D_{pm} \cup D_T$. For the Error Correction Data, we use GPT-4 as the M_T, which is one of the most powerful closed-source LLM. We first generate $n = 100$ Error Correction Data and evaluate the quality by humans. We found that 98% of them give the right correction about the user's intention and 85% of them give the right correction about transforming the user's needs into correct drawing prompts. Therefore, we consider the quality of the error correction to be acceptable for further training. During inference, given a well-trained model M_w, we apply the same prompt P_I as training. That is, $P_I = P_T$.

Two-Step Inference To fully explore our model's ability, we additionally introduce a two-step inference procedure given that our model is fine-tuned with error correction data to be blessed with the ability to conduct self-correction. To be more specifically, given the query of the ith test example $q_t^{(i)}$, we first use P_i to get a response $r_1^{(i)}$, where

$$r_1^{(i)} \sim M_w(P_I \oplus q_t^{(i)}) \tag{3}$$

Afterwards, $r_1^{(i)}$ is sent into the model again with $q_t^{(i)}$ together with the correction prompt P_c and get the corresponding correction $EC^{(i)}$ and suitable output $r_2^{(i)}$. We choose $r_2^{(i)}$ to be the final output.

$$EC^{(i)}, r_2^{(i)} \sim M_w(P_c \oplus q_t^{(i)} \oplus r_1^{(i)}) \tag{4}$$

5.3 Evaluation

Modality Switching. In Table 2 we include the

Table 2: Benchmarking on Modality Switching ability of different models, with T and I represents text and image respectively. On the left of → is the input modality, while on the right of → is the output modality. The accuracy is calcualted by Equation.3.2.1. The best result is emphasized in **bold**.

Model	**Round1**				**Round2**				**Round3**				**Avg.**
	$T \to T$	$T \to I$	$I{+}T \to I$	$I{+}T \to T$	$T \to T$	$T \to I$	$I{+}T \to I$	$I{+}T \to T$	$T \to T$	$T \to I$	$I{+}T \to I$	$I{+}T \to T$	
Qwen-VL-0-shot	92.46	7.42	1.11	**100**	76.75	34.13	2.54	**100**	78.16	27.98	4.24	99.58	52.03
Qwen-VL-1-shot	93.90	8.42	1.04	**100**	79.78	34.27	2.29	**100**	79.48	27.58	3.98	**100**	52.56
NExT-GPT	74.89	59.37	58.88	90.43	89.66	38.95	31.35	91.21	96.06	16.87	20.49	93.75	63.49
SEED-LLaMA	94.71	89.69	92.30	97.18	98.79	84.82	85.98	98.13	**99.01**	74.77	76.74	97.28	90.78
DialogGen-A	80.00	**97.66**	97.04	95.25	65.56	98.40	93.64	88.75	64.93	**98.77**	93.28	76.85	87.51
DialogGen-B	85.41	98.44	98.15	97.87	90.47	**99.20**	**99.57**	94.58	93.93	100	99.00	91.00	95.63
DialogGen	**94.97**	96.87	**98.88**	97.82	**97.04**	98.79	**99.00**	95.81	95.19	97.53	98.23	95.83	**97.24**

comparison result on Modality Switching accuracy of different models. Note that both Qwen-VL-0-shot and Qwen-VL-1-shot have very poor performance in identifying the user's drawing intention, which is reflected in the small number in $I{+}T \to I$. We also check the inference output for $I{+}T \to T$ and find that these two models use texts to reject to answer the question related to Visual-Question-Answering in most of the cases, leading to the intention being recognized as correct. This contributes to their near-perfect performance in scenario $I{+}T \to T$.

As can be seen in the table, for DialogGen-A, the model tends to recognize the user's intention as draw rather than talk and hence has relatively high accuracy when the output modality is image. For NExT-GPT, it has higher accuracy in the case when the output modality is text and severely worse performance when the output modality is image. This is because it has four output modalities. When the ground truth modality is image there are often hints in the query to inform the model to produce content other than text, which somehow lets the model produce audio or video instead. And as the conversations go deeper, both the two models show several performance degradations in the scenario that they are not good at.

For experiments including error correction (Table 1), we can see that by incorporating Error Correction data, both models can have obvious performance improvement, especially when the diversity of the training dataset is relatively small (DialogGen-A), the average success rate of switching modalities increases from 88.4 to 95.3. Besides, incorporating error correction data helps improve the stability of the model as the conversations go deeper. As for the two-step inference, the benefit is less obvious.

For comparative experiments including bilingual training (Table 6). To better illustrate the effectiveness of adding bilingual data, we eliminate all the training except DialogBen-train and split it into English (DialogBen-train-en) and Chinese (DialogBen-train-cn). We can figure out that both the model trained in English and the model trained in Chinese can benefit from adding the same training data except for the language difference.

Table 3: Comparison of Cohence VQA Score on different models. $^{\uparrow}$ indicates higher is better, **bold** indicates the best result. DialogGen-HunyuanDiT has the best performance and the Coherence VQA Score perfectly aligns with Human evaluation.

Model	Coherence VQA Score$^{\uparrow}$	Human$^{\uparrow}$
NextGPT-SD-v1-5	0.5153	0.5524
DialogGen-A-SD-v1-5	0.6371	0.6732
DialogGen-SD-v1-5	0.6403	0.6911
SEED-LLaMA-SD-v2-1	0.5776	0.6313
DialogGen-SD-v2-1	0.6468	0.7277
DialogGen-HunyuanDiT	**0.6514**	**0.7559**

Coherence VQA. Note that our DialogGen builds a multi-modal interactive dialogue system that connects with users and T2I models and hence is compatible with any T2I models. We provide the coherence VQA score and the human evaluation[1] of different Multi-modal Interactive Dialogue systems.(See Table 3) Here DialogGen-HunyuanDiT is DialogGen equipped with HunyuanDiT and DialogGen-SD-v1-5 and DialogGen-SD-v2-1 is DialogGen equipped with Stable Diffusion (Rombach et al., 2022) v1.5 and v2.1 respectively. They are the same T2I models equipped with NExT-GPT and SEED-LLaMA.

In Fig 5 in the appendix we also provide the visualization of the Coherence VQA Score of different models. We refer the readers to the analysis in the appendix. Besides, given that the coherence VQA score perfectly aligns with human evaluation, we consider the coherence VQA score to be a suitable proxy that reflects what extent the output image satisfies the user's need.

[1]Details can be found in Appendix.

Qualitative Examples. We provide more qualitative examples and demo conversations in Fig. 6, 7, 8 in Appendix and refer the readers to the appendix.

6 Conclusions

In this work, we propose DialogGen, a Multi-modal Interactive Dialogue system (MIDS) for Multi-turn Text-to-Image generation, which is trained with carefully curated instruction tuning and Error Correction data generated by more powerful LLMs. In addition to that, in response to the need for a comprehensive benchmark to evaluate rapidly emerging new MIDS, we propose DialogBen. It consists of 9957 three-turn multi-modal conversations together with two evaluation metrics, namely modality switching accuracy and coherence VQA score to comprehensively assess the ability of a Multi-modal Interactive Dialogue System.

Future work. Current training data collection is aligned with the T2I model's characteristics. It can be designed to be aligned with human preference in the future and hence we can train the MLLMs to satisfy human preference using algorithms like (Rafailov et al., 2024)

7 Limitations

The limitation lies in the resource requirement when conducting the re-captioning for the training data of the T2I model to conduct the drawing prompt alignment. However, fine-tuning open-sourced T2I models with a small subset of high-quality data with known distribution P_{hq} has been widely adopted. In this case, we can simply sample $(I_{icl}, T_{icl}) \sim P_{hq}(I, T)$ to simplify the procedure. Besides, the potential risk of DialogGen lies in the lack of content filtering of the pre-trained MLLM and T2I models.

8 Acknowledgements

Minbin Huang and Hong Cheng are supported by project #MMT-p2-23 of the Shun Hing Institute of Advanced Engineering at The Chinese University of Hong Kong, and by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14217622).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023b. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. 2024. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2023a. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*.

Lai Zeqiang, Zhu Xizhou, Dai Jifeng, Qiao Yu, and Wang Wenhai. 2023. Mini-dalle3: Interactive text to image by prompting large language models. *arXiv preprint arXiv:2310.07653*.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Dataset Details

Table 4 lists the detailed data composition of the benchmark. As stated in Sec.5, our training datatset consists of D_o, D_p, D_{pm} and DialogGen-train. In Tab. 4 we summarize the statistics of the components of the whole training datasets and in Example 2 and Example 3 we give some examples of the instruction tuning data we used during our training.

B Coherence VQA

We provide visualization of Coherence VQA Score in Fig. 5. Here DialogGen-X is DialogGen equipped with Hunyuan-DiT, and DialogGen-SD is the one equipped with Stable Diffusion v1.5. NExt-GPT (Wu et al., 2023) and SEED-LLaMA (Ge et al., 2023) are two comparative models in the setting of Multi-modal Interactive Dialogue System.

As can be seen in the figure, in the first turn of the dialogue, both DialogGen-X and DialogGen-SD generate correct images except for a flaw in the action **jumping** for DialogGen-SD. While both NExT-GPT and SEED-LLaMA make two mistakes and the coherence VQA score correctly capture the mistakes. In the second turn of the dialogue, DialogGen-SD misses the action **jumping** and both NExT-GPT and SEED-LLaMA make two mistakes. Looking more closely we can find the DialogGen-X have a relatively better consistent main body. Therefore, we consider coherence VQA score to be a suitable proxy that reflect to what extend the output image satisfy the user's need.

Moreover, we also provide human evaluation on a small subset of $k = 30$ examples of whether all the models in Tab 3 successfully identify the correct intention of the user as drawing and generating the corresponding images. For each of examples i with corresponding n_i questions, we ask the human annotators to give scores $\{s_{ij}\}_{j=1}^{n_i}$ to measure how likely the answer to these questions are true, from $s_{ij} = 5$ being the most likely to $s_{ij} = 1$ being the least likely. Afterwards, we take the average of these scores to be the final human evaluation score $s_h = \sum_{ij} s_{ij} / \sum_{i} n_i$. Finally, we scale it into the interval between 0 and 1 to have a more intuitive comparison between coherence VQA.

Example 4: Correction examples

History: (empty)
Question: I'd like to know more about rockets. Could you draw me a picture of rockets?
Original Output:`<draw>`A rocket is a large, powerful, and complex machine that is used to transport people and goods into space. It is usually made of steel and aluminum and is equipped with engines, fuel tanks, guidance systems, and other equipment. The picture shows a large rocket with a long tail, flying high in the sky.
Correction: The output violates rule 3. The assistant's description misses the main points of the asking for a visual image of a rocket.
Correct Solution:`<draw>`A rocket propelled upward by burning flames is moving through space, the Milky Way and stars in the background, the shot is panoramic, and the style is cartoonish.

C More visualization results

We present some qualitative examples of our model and NExT-GPT in Fig. 6. All the Q1 and Q2 are coherent conversations. The content within the bracket is transformed drawing prompt for the given scenario. Our model generates correct drawing prompts in response to the user's demand in all these cases. Noticeably, on the top right R2 the correct prompt failed to let Stable Diffusion generate the correct style while ours (X) succeeded. This is because our DialogBen-train is collected so that it has already been aligned with X. We also provide some demo conversations in Fig. 7 and Fig. 8. DialogGen-HunyuanDiT can accurately respond to the users' need and perform coherent multi-turn image generation.

Example 2: Instruction Tuning Data

D_o:

user: 

What is hanging from the traffic light pole?
assistant: A green street sign is hanging from the traffic light pole.

D_p:
user: I want to see a squirrel in the forest.
assistant: <draw>A lively squirrel jumping in the forest, the background is dense forest, the lens is close-up, the style is cartoon style.

D_{pm}:

user: 
What is the cat doing in the image?
assistant: The cat is coming out from some curtains onto the couch and is sitting or standing on top of it.
user: Design a logo that reflects environmentalism.
assistant: <draw>A circular logo, the main body is a hand in the care of a young sapling, the background is light green symbol of environmental protection and peace, viewing angle for squint, showing the harmony between human and nature, the style is simple style.

Example 3: Instruction Tuning Data

DialogGen-train:
Sample 1:
user: I love fast trains. Can you draw me a picture?
assistant: <draw>A black steam train gallops along the tracks. The background is a forest in fog. The lens is close-up. The style is realistic.
user: Draw the train in cartoon style.
assistant: <draw>A black steam train gallops along the railway tracks. The background is a forest in fog. The camera is close-up. The style is cartoon style.
user: I prefer smooth lines, so try the line drawing style.
assistant: <draw>A black steam train gallops along the tracks. The background is a forest in fog. The lens is close-up. The style is line drawing.

Sample 2:
user: What is the woman in the picture doing?
assistant: The woman in the picture is practicing dance.
user: Can you add a little girl watching a girl dance to this picture and add some interactive elements?
assistant: <draw>A woman dressed in a dance outfit was practicing dance in the room. Beside her sat a little girl with wide eyes, as if she was watching intently. The background was a simple and bright dance practice room. The camera was in the middle ground, and the style was realistic.
user: How does the little girl feel, how does she feel about the dance?
assistant: The little girl loved this dance and envied the dancing women. She seemed to be imagining that she could dance so beautifully in the future.

Round1 (%)				Round2 (%)				Round3 (%)			
$T \to T$	$T \to I$	I+$T \to T$	I+$T \to I$	$T \to T$	$T \to I$	I+$T \to T$	I+$T \to I$	$T \to T$	$T \to I$	I+$T \to T$	I+$T \to I$
24.6	28.0	23.7	23.7	24.1	27.6	23.3	25.0	24.2	23.9	24.5	27.2

Table 4: The data proportion for each modality switching scenario in each round. The input modalities lie before $\to$ and output modalities lie after, where I,T represent image and text respectively.

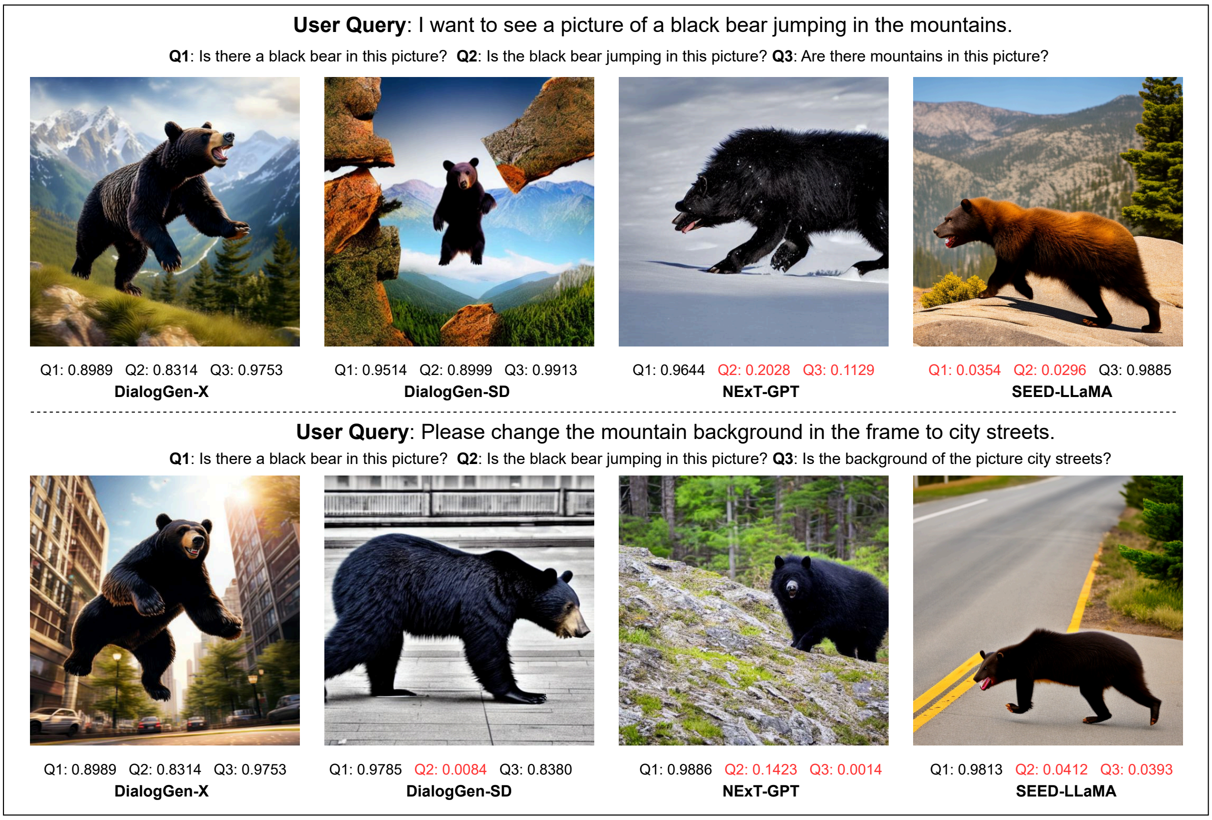


Figure 5: Visualization of Coherence VQA on different models. The numbers after each question are the probability of the model answering correctly about the corresponding question. Red numbers indicates the wrong answers.

Data	Descriptions	Size
D_o	It contains open-sourced uni-modal single-turn conversations like alpaca (Taori et al., 2023) and multi-modal multi-turn conversations like LLaVA (Chen et al., 2023b)	8.8K
D_p	It contains carefully generated single-turn text to drawing prompts conversations.	3.6K
D_{pm}	It contains randomly concatenated samples from D_o and D_p to form multi-turn conversations that doesn't necessarily have the coherent semantics.	5K
DialogBen-train	It contains carefully generated three-turn semantically coherent conversations. The input involve texts and images and the out involves texts and images (transformed drawing prompts)	4.5K

Table 5: Statistics of the instruction tuning datasets of DialogGen

Table 6: Comparison of adding bilingual training on Modality Switching (MS) Accuracy (%) of models under different training data. *cn* and *en* represent Chinese and English data respectively.

Model	Training	Evaluation Data	DialogBen-MS			
			Round1	Round2	Round3	Avg
Qwen-VL	DialogBen-train-*en*	DialogBen-test-*en*	77.8 / 97.6 / 97.0 / 95.6	77.5 / 98.4 / 96.2 / 88.7	79.9 / 95.6 / 95.7 / 79.6	90.0
	+ DialogBen-train-*cn*	DialogBen-test-*en*	78.2 / 96.9 / 97.0 / 96.0	83.0 / 97.6 / 96.6 / 89.5	82.5 / 96.7 / 95.4 / 85.4	91.3 (+1.3)
Qwen-VL	DialogBen-train-*cn*	DialogBen-test-*cn*	82.5 / 98.0 / 97.0 / 97.0	76.0 / 99.0 / 94.0 /92.0	71.0 / 97.9 / 93.6 / 83.8	90.1
	+ DialogBen-train-*en*	DialogBen-test-*cn*	87.0 / 98.4 / 97.7 / 98.7	90.4 / 99.2 / 95.3 / 95.0	86.0 / 98.0 / 92.6 / 89.2	94.0(+3.9)

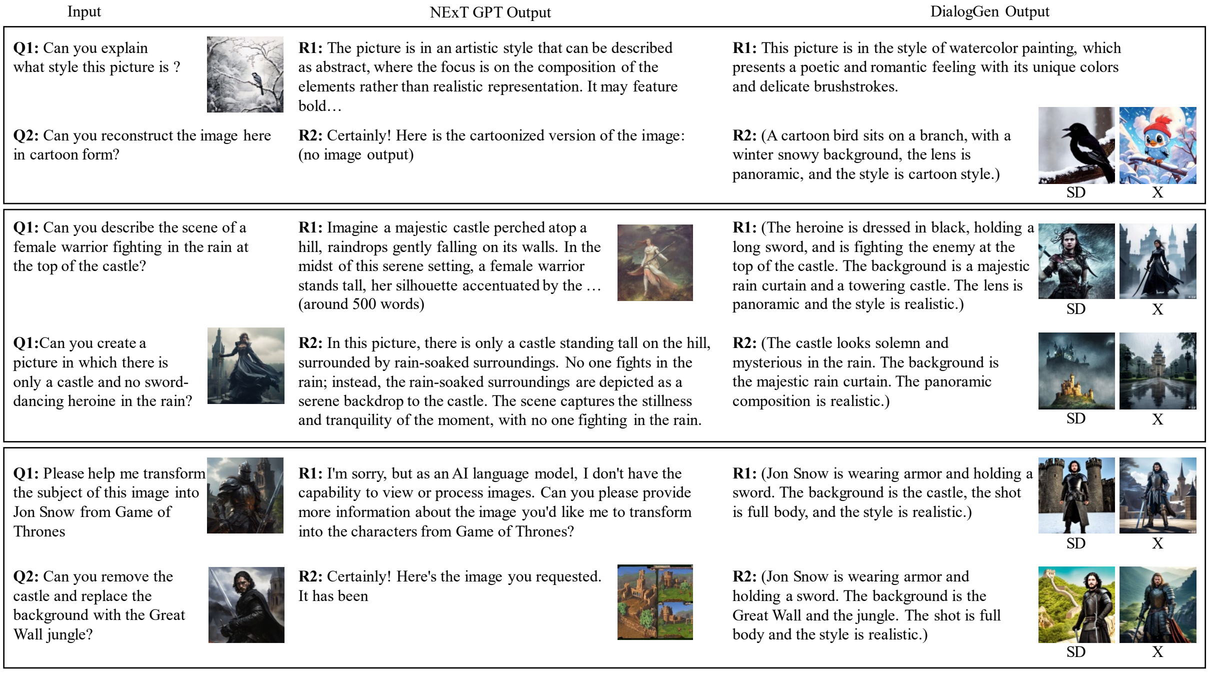


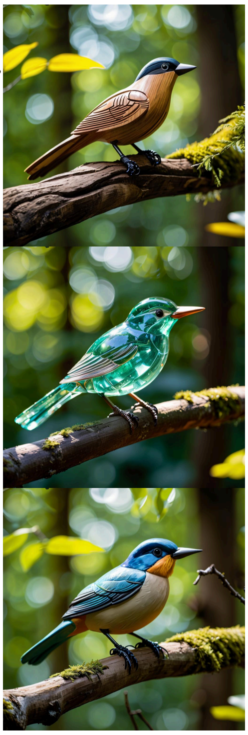
Figure 6: Visualization of output results of NexTGPT, DialogGen-SD, and DialogGen-Hunyuan on the DialogBen benchmark. DialogGen has better performance in generating output of correct modality and higher semantic coherence.

Figure 7: Visualization of inference result of DialogGen-HunyuanDiT

#Turn 1:
cn: *请你画一只木制的鸟放在一个枝头上。*
en: *Please draw a wooden bird and place it on a branch.*

#Turn 2:
cn: *把鸟的材质改变成玻璃*
en: *Change the bird's material to glass*

#Turn 3:
cn: *现在把画面的鸟更换成陶瓷质地。*
en: *Now replace the bird in the picture with a ceramic texture.*



#Turn 1:
cn: *请画一位坐在公园长椅上的老人，他在喂鸽子，天空下着细雨。*
en: *Please draw an old man sitting on a park bench, feeding the pigeons, and it is drizzling.*

#Turn 2:
cn: *我想改变人物的年龄，将图画中的老人变为青年，其余内容保持不变。*
en: *I want to change the age of the characters, turning the old man into a young man in the drawing, and keep the rest of the content the same.*

#Turn 3:
cn: *将图画中的人物年龄再调整一次，这次，请将青年变为儿童。*
en: *Adjust the age of the characters in the picture again. This time, please change the youth to a child.*

Figure 8: Visualization of inference result of DialogGen-HunyuanDiT