Mini DALL·E 3: Interactive Text to Image Generation by Prompting Large Language Models

Zeqiang Lai [1] Xizhou Zhu [23] Jifeng Dai [23] Yu Qiao [2] Wenhai Wang [4]
[1]Beijing Institute of Technology [2]Shanghai AI Laboratory [3]Tsinghua University [4]CUHK
https://minidalle3.github.io/

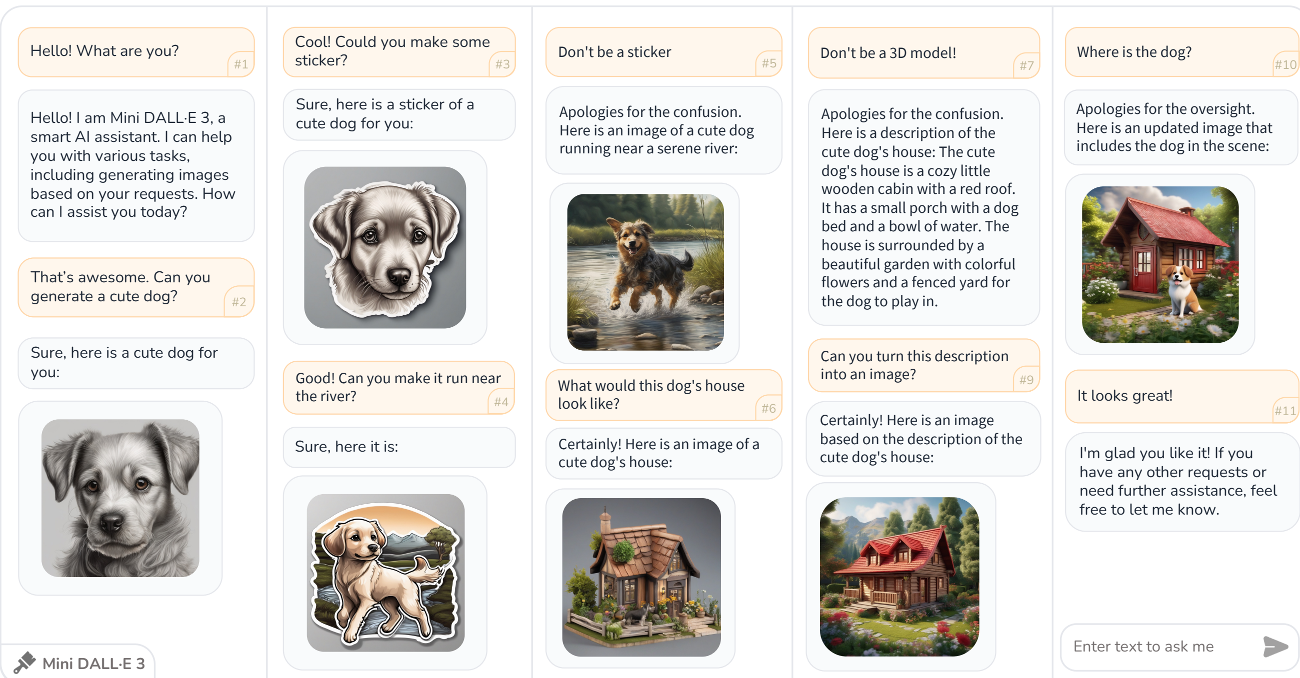

Figure 1. Examples of two interactive text-to-image conversations produced by Mini DALL·E 3. In these cases, people can ask the agent to generate images via natural language and request an edit if the results are unsatisfactory. The generation and editing can be completed in a multi-turn dialog with recognition of the conservation context.

Abstract

The revolution of artificial intelligence content generation has been rapidly accelerated with the booming text-to-image (T2I) diffusion models. Within just two years of development, it was unprecedentedly of high-quality, diversity, and creativity that the state-of-the-art models could generate. However, a prevalent limitation persists in the effective communication with these popular T2I models, such as Stable Diffusion, using natural language descriptions. This typically makes an engaging image hard to obtain without expertise in prompt engineering with complex word compositions, magic tags, and annotations.

Inspired by the recently released DALL·E 3– a T2I model directly built-in ChatGPT that talks human language, we revisit the existing T2I systems endeavoring to align human intent and introduce a new task - ***interactive text to image*** *(iT2I), where people can interact with LLM for interleaved high-quality image generation/edit/refinement and question answering with stronger images and text correspondences using natural language. In addressing the iT2I problem, we present a simple approach that augments LLMs for iT2I with prompting techniques and off-the-shelf T2I models. We evaluate our approach for iT2I in a variety of common-used scenarios under different LLMs,* e.g., *ChatGPT, LLAMA, Baichuan, and InternLM. We demonstrate that our approach could be a convenient and low-cost way to introduce the iT2I ability for any existing LLMs and any text-to-image models without any training while bringing*

Preliminary version. Work in Progress.

little degradation on LLMs' inherent capabilities in, e.g., question answering and code generation. We hope this work could draw broader attention and provide inspiration for boosting user experience in human-machine interactions alongside the image quality of the next-generation T2I systems.

1. Introduction

The evolution of artificial intelligence content generation has been significantly accelerated by the proliferation of text-to-image (T2I) diffusion models [18, 20, 41, 43]. Within just two years of rapid development since 2021, it was unprecedentedly of high quality, diversity, and creativity that the state-of-the-art T2I models [4, 13, 39–41, 43, 55] could generate. For the first time, "talk to paint" is no longer a daydream, and complex surrealistic arts can be generated via textual descriptions, with stronger expressive ability than previous unconditional and class conditional image generation systems as shown in Fig. 2.

However, it is unfortunate that most of the existing T2I models, such as Stable Diffusion [41], are still limited in understanding natural language. In other words, people have to learn to write complex text prompts to obtain the best results, which fit the used models but are not necessarily user-friendly and straightforward for humans, as illustrated by Fig. 6. As a result, this typically makes an engaging image hard to obtain without expertise in prompt engineering with proper word compositions and sometimes weird phrase organizations. Besides, there have been dozens of different textual and numerical configurations in a diffusion-based T2I pipeline, such as CFG scale, word weighting, negative prompts, and style keywords, which are also complicated for non-professional users.

To make it easier for users to utilize T2I models, Stable Diffusion (SD) WebUI [2] is first created to provide a user-friendly web UI to access the latest techniques without any coding. However, a typical workflow of generating a satisfactory image usually involves several stages, *e.g.*, generation, variation, super-resolution, *etc*. This makes the tab-based interface of SD-WebUI somewhat awkward to use. Therefore, ComfyUI[1] was designed by utilizing a graph/nodes interface that connected different stages via nodes and edges, which makes workflows more clear. Nevertheless, these software tools still could not solve the problem of complicated configurations required for a charming image. This urges the development of Fooocus[2] – a tool with a bunch of built-in optimizations and quality improvements. Fooocus frees users from complex parameter-tuning, but it still requires them to write a proper and precise text prompt for the desired images. However, this can be challenging in some cases, such as when the required scenes are artistic conceptions rather than specific objects, or when the users have no idea how to describe what they want to generate, etc.

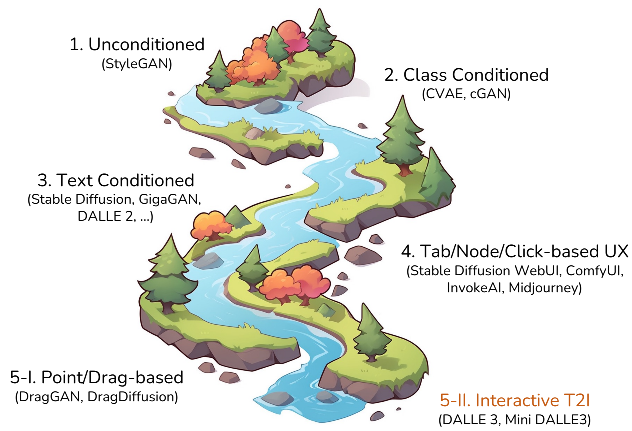

Figure 2. The evolution of image generation systems.

Generally, it might be difficult for users to come up with the right prompts and configurations at once, but it is much easier to tell what they want or do not want via natural language if the first version is unsatisfactory, *e.g.*, "Don't be a sticker" and "Where is the dog?", as shown in Fig. 1. Moreover, it would be more straightforward to perform a multi-turn conversation with T2I models to iterate the images over and over again, mimicking the communication processes between human designers and their customers. These analyses reveal a promising direction for building the next generation of T2I systems with a new human-machine interface using natural language – a system that is able to infer users' intentions and automatically generate the proper text prompts leveraging the reasoning abilities of large language models (LLM). This is not only because natural language is the easiest way that everyone can master, but also because it frees users from brainstorming sophisticated textual descriptions and requires only simple instructions instead (see Fig. 6 for more illustrations).

Inspired by the recently released demo of DALL·E 3 [35] – a powerful T2I model directly built-in ChatGPT that utilizes human language, we revisit existing techniques aimed at aligning human intent in generating images and introduce a new task called **interactive text to image** (iT2I). This task is featured by several aspects, including 1) *Multi-Turn*: users are allowed to chat with the system (typically powered by LLMs) to progressively specify requirements, shortcomings, and suggestions of the expected/generated images; 2) *Consistency*: the ability to keep identity for consistent multi-turn image editing, series characters creation, *etc*.; 3) *Composability*: the ability to be composed with/built-in ex-

[1] https://github.com/comfyanonymous/ComfyUI
[2] https://github.com/lllyasviel/Fooocus

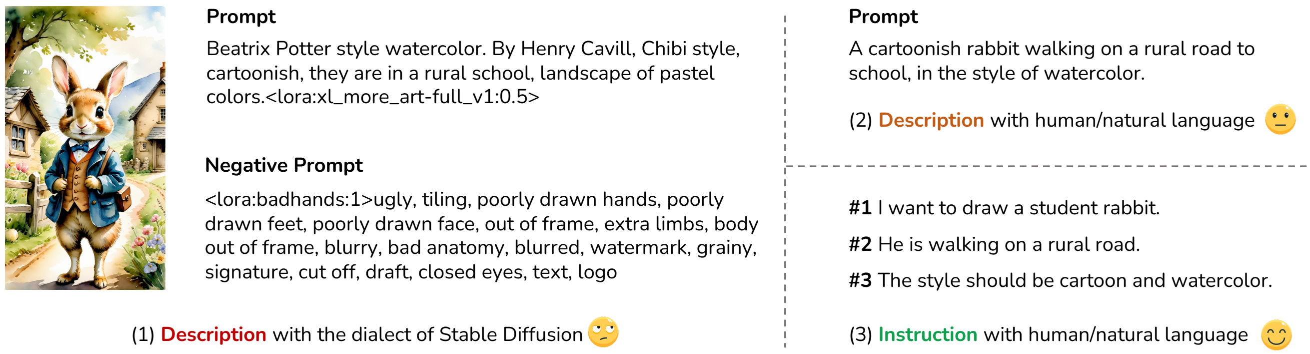

Figure 3. Illustrations of different human-machine interfaces for T2I systems.

isting chat assistants for interleaved image generation and (visual) question answering for seamless user experience. All these properties make iT2I systems powerful tools for a wide range of applications, from content generation and design to interactive storytelling and more.

As an initial solution to address this problem, we propose a simple yet effective approach that enhances language models for iT2I using prompting techniques and pretrained text-to-image models. Specifically, we prompt the LLM to instruct it to generate an image with an intermediate textual description enclosed by special tags. After detecting the special tags, the description is parsed and transformed through a prompt refinement module. Then, a pretrained T2I model is employed to generate the image. We evaluate our approach across various common use cases and different language models such as ChatGPT [7, 36], LLAMA [48], Baichuan [56] and InternLM [46]. Our results demonstrate that our approach can easily enable iT2I capabilities in any existing language model and text-to-image model without the need for additional training. Furthermore, it has minimal impact on the language models' inherent abilities in question answering and code generation.

We hope this work could draw broader attention and provide inspiration for boosting user experience in human-machine interactions alongside the image quality of the next-generation T2I models.

2. Related Works

Text-to-Image Generation. Text-to-image (T2I) generation is a widely-explored research area at the intersection of computer vision and natural language processing. Notable approaches include generative models, like Variational Autoencoders (VAE) [22, 47], Generative Adversarial Networks (GAN) [17, 21], and autoregressive models [12], which enable image synthesis guided by textual descriptions. Recent multimodal models like CLIP [38] and DALL·E [39] have further improved alignment between text and generated images, while the birth and development of diffusion models [4, 13, 40, 41, 43, 55] have pushed the boundaries of text-image interactions.

Image Generation Interface. There are a variety of different approaches for image generation and editing – each possesses its own merits and drawbacks. The most straightforward ones are text-based approaches where people write text prompts for either image generations [40, 41] or image editing [6, 61]. Besides, image-based approaches are also popular. In this case, people either provide a reference image asking the T2I models to generate image variations [40, 59], or provide edge/depth maps to control the image layout [27, 34, 64], or performing image translation with a style image [1, 45], or asking generating images of a given subject [25, 57]. To facilitate the precise control, point-based approaches [31, 49] are widely adopted by utilizing state-of-the-art localization methods [23, 30]. Recently, drag-based approaches [11, 28, 29, 33, 37, 44, 62] are also proposed for more interactive experience. As for UX design, there are Rich-T2I [15] and DialogPaint [53], which share similar spirits as ours. In the literatures of integrating T2I to LLM, there are NExT-GPT [54], GILL [24], DreamLLM [9], SEED [16]. Although these methods also provide the capability for interleaved text-image generation, they are not specifically designed for iT2I mostly and are limited to image quality and multi-turn correspondence.

Prompting LLMs. The in-context-learning capability [7] is one of the strongest advantages of LLMs. It enables users to freely customize LLMs for a particular task or enhance the capabilities of LLMs by simple prompting. For example, chain-of-thoughts [52] is the first prompting technique that enhances LLMs by asking them to generate a series of intermediate reasoning steps. After that, there are also a number of improved prompting techniques that leverage the heuristic of majority voting [51], backtracking [58], and graph of thoughts [5]. In this work, we also provide an ap-

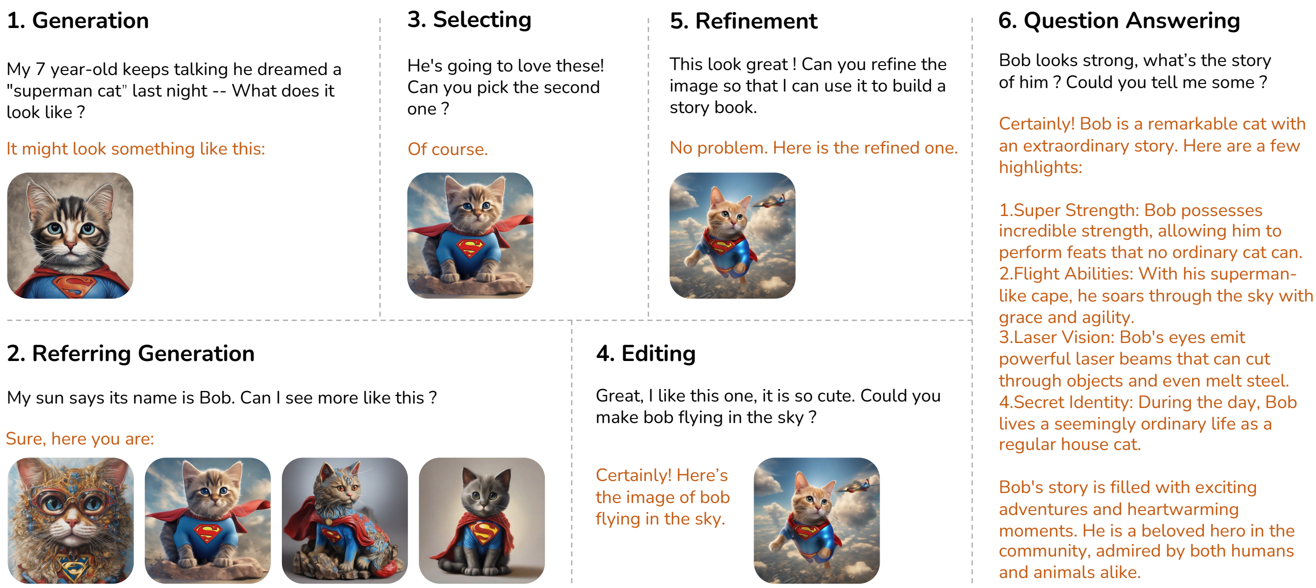

Figure 4. Illustration of 6 types of interactions in interactive text-to-image workflow.

proach to augment LLM with iT2I ability via prompting, as it can be rapidly applied to any existing LLMs without any training.

3. Interactive Text to Image

Interactive Text to Image (iT2I) aims to provide a user-friendly approach to generate images that meet user requirements in an interactive manner. Users can instantiate a multi-turn dialogue between humans and AI agents, where they can communicate requirements, shortcomings, and suggestions of the generated images or the expected ones with natural language.

3.1. Problem Definition

Precisely, the iT2I problem can be defined as the task of generating images from textual descriptions in a way that the generated images closely align with the provided text, ensuring that the generated visual content accurately represents the textual information. There are some notable properties of iT2I systems:

Multi-Turn refers to the ability of the system to engage in a dynamic and iterative dialogue with the user. Unlike traditional text-to-image systems that may generate a single image based on a static textual input, multi-turn iT2I systems can accept multiple rounds of textual input, enabling users to refine and specify their visual requirements through an ongoing conversation. This property enhances the user experience and allows for more fine-grained control over the generated images.

Consistency means that these systems can automatically determine if they should take into account not only the current textual input but also the previous visual context. It involves persisting the visual identity of images in different rounds of generations. This capability enables iT2I systems to perform consistent multi-turn image editing/refinement, produce personalized and contextually relevant objects/characters, *etc*.

Composability relates to the ability to combine or integrate image generation with other tasks. This means that the ability of image generation should be modular and compatible with the inherent abilities of LLMs, allowing users to seamlessly incorporate them to perform interleaved conversations for querying both text and visual content.

3.2. Types of Instruction

As shown in Fig. 4, there are various instructions that could be found in an iT2I system, such as generation, editing, selecting, and refinement. Different instructions could have varying levels of complexity when it comes to interpretation. Some instructions can be effectively addressed by leveraging the capabilities of an LLM, such as selecting, which primarily involves textual decision-making. However, certain instructions may necessitate a deeper synergy between the LLM and the T2I models.

Generation refers to the process of generating entirely new images based on a given textual description. In this context, the iT2I system creates images or illustrations from scratch, attempting to capture the essence and details of the provided textual input. It essentially transforms queries into neural representations or prompts for T2I models.

Referring generation is another variant of generation, where the system generates images that refer to or are inspired by existing objects, scenes, or concepts mentioned in the textual input and appear in the context.

Selecting is a relatively straightforward instruction that involves choosing or picking from a set of pre-existing or bag of generated images based on the textual input.

Editing performs the task of modifying or refining existing images in response to textual instructions. This may involve altering specific attributes of an image, enhancing or diminishing certain features, or adapting the image to match the requirements outlined in the instruction.

Refinement means to further enhance or optimize an existing image to better align with the textual description. While editing involves making specific modifications, refinement often involves fine-tuning the visual output to achieve a higher level of detail, realism, or accuracy in accordance with the provided textual guidance.

Question Answering is the inherent ability of LLMs. An iT2I system should be able to persist the ability as much as possible, as it is crucial to provide a coherent experience interleaving images and text for users.

3.3. Discussion

In the literature of image editing and multi-modal LLM, there are a number of works that are closely related to iT2I. Most of these related works could provide interactive interfaces. For example, InstructPix2Pix [6] and its follow-up works [63, 65] could be repeatedly applied to a single image to achieve multi-turn image editing. However, these interactive multi-turn abilities only apply to image editing instructions. There are also multi-modal LLMs [9, 16, 24, 54] that could generate response with interleaved text and images, but most of them focus more on (visual) question answering with multi-modal responses rather than interactive image generation. The key vision of iT2I is to build a chat-based system that could respond to all image generation/editing instructions in a multi-turn, consistent, and composable manner. This is the major difference between iT2I from all previous works/tasks.

4. Mini-DALLE3

In this section, we depict a blueprint of an iT2I system, which we refer to as Mini-DALLE3. The overall architecture of Mini-DALLE3 is illustrated in Fig. 5, and it comprises several key components: an LLM, a router, an adapter, and T2I models. The LLM can be an existing text-only LLM, such as ChatGPT [36] and LLaMA [48], or multi-modal LLM [50]. It is responsible to analyze user intentions and produce the proper outputs in text or neural representations. The router would automatically dispatch the parsed image representations (if there exist ones in the LLM output) to the image generation module. The adapter transforms the image representations to better fit the backend T2I models. Depending on the type of image representations, the adapter can be a neural network if the image representations are neural embedding or prompt refinement modules with handcrafted rules or LLM. Next, we illustrate a simple yet effective instantiation of Mini-DALLE3 architecture by prompting large language models.

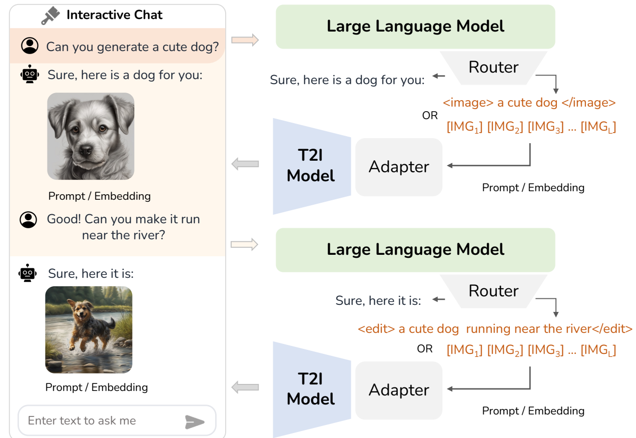

Figure 5. **Pipeline Overview.** Mini-DALLE3 consists of two stages, with 1) a router that analyzes the response from the prompted/finetuned LLM and dispatches the demand for image generation if needed, and 2) an adapter that transforms the image embedding or descriptions for subsequent T2I models.

4.1. Multi-Turn Interaction by Prompting LLM

Multi-turn interaction lies at the heart of interactive text-to-image. It possesses the requirements of integrating textual/visual context and understanding instructional instead of descriptive messages. To address it, we propose to leverage the stronger context-understanding ability of LLMs by prompting them to pretend to generate images via textual descriptions. This intermediate textual description not only provides stronger flexibilities to augment the system capabilities with plug-and-play modules such as prompt variation/refinement but also enables us to utilize numerous pretrained LLMs and T2I models without heavy finetuning.

Image Generation as Function Call. Specifically, we utilize the few-shot prompt as shown in Fig. 6 to transform the problem of multi-turn image generation into a problem of multi-turn textual description generation. Our prompt entails several key steps. Initially, we define the LM's role and explicitly convey to it that it possesses the ability to generate images. Subsequently, we request the LM to produce images by generating descriptive text enclosed within ⟨image⟩ tags. If the generated images exhibit a high degree of correlation with previous ones, the LM is instructed to generate "edit" rather than generate "image". Finally, we provide a few number of few-shot examples to further guide the LM's responses. Leveraging the robust in-context learning capabilities inherent in advanced LLMs, we observe that

```
You are Mini-DALLE3, a smart AI assisant that could generate image if
the user request asks.
You should include the generated image by its description with the format
of " <image> HERE IS THE DESCRIPTION </image> "
If the generated image is strongly correlated with previous image, you
should use <edit> </edit> tag instead of <image> </image>.

Here is an example

User: My 5 year-old keeps talking about a "super-duper sunflower
hedgehog" -- What does it look like ?
AI: It might looks like as <image> a super-duper sunflower hedgehog
</image>
User: My daughter says its name is Larry. Can I see more like this ?
AI: Sure, here it is  <edit> a super-duper sunflower hedgehog </edit>
User: She's going to love these! Can you show me Larry's house ?
AI: Of course, <edit> a super-duper sunflower hedgehog, standing in
front of a house </edit>
User: Larry is cute, what makes him so super-duper ?
AI: Larry is super-duper for a myriad of reasons! Here's why: 1. Sunflower
Petals as Quills: Unlike other hedgehogs, Larry boasts sunflower petals
as quills. This not only makes him stand out, but it also gives him a bright
and cheerful demeanor. 2. Kind Hearted: Larry is known throughout the
meadow for kindness.
User: Awwww...can you show me Larry being "kind hearted"
AI: <edit> a super-duper sunflower hedgehog, kind hearted </edit>
User: Can your design some stickers ?
AI: Sure, <edit> a sticker of a super-duper sunflower hedgehog </edit>

Here is another example

User: can you generate a dog ?
AI: Sure, <image> a cute dog </image>
User: can you generate a cat ?
AI: Sure, <image> a cute cat </image>
```

Figure 6. The few-shot prompt for iT2I generation.

this approach yields favorable outcomes. The LM successfully generates images accompanied by coherent textual responses, as illustrated in Fig. 1. Importantly, these capabilities can be harnessed without the need for specialized training and can be swiftly integrated into existing LLMs.

Prompt Refinement&&Variations. Although we can generate textual descriptions that integrate the information from context by prompting LLMs, the descriptions might not be sufficient to generate high-quality images. Therefore, we propose to leverage another round of prompt refinement to transform the vanilla descriptions to better fit subsequent T2I models. It is worth noting that the prompt refinement can also apply to embedding if the previous intermediate representation is embedding. In this instantiation, we perform text transformation by prompting LLM again with the following few-shot prompt. Furthermore, we can perform prompt variation by repeatedly performing different prompt refinements, which is useful for responding to the request to generate a list of images.

4.2. Hierarchical Content Consistency Control

Content consistency is another important aspect of an iT2I system. Although similar topics (subject-driven T2I, example-driven T2I, personalization, concept learning) are widely explored in the context of conventional T2I [26, 42, 57], only a few works explore the multi-turn scenarios and seldom works explore the integration of these abilities into a single unified system. Our decomposition makes it possible to utilize existing T2I models that were not designed for multi-turn scenarios. For example, the edited description of Prompt-to-prompt [32] can be automatically generated through LLM in an interactive manner.

Specifically, we leverage the off-the-shelf T2I models that take previous images as additional input to ensure consistent multi-turn generation. To better ensure the image quality, we introduce a hierarchical control strategy that utilizes different models for different levels of content changes. For small content changes that can be described in a few words, such as changing styles, word weighting, and simple object manipulation, we adopt the models of Prompt to prompt [32] and MasaCtrl [8]. We utilize IP-Adapter [60] to perform large content changes as these models are more flexible for the input textual prompts.

Task	GPT3.5-Turbo Original	GPT3.5-Turbo Mini-DALLE3
Abstract Algebra	42.42	43.43
High School Physics	40.00	38.67
Marketing	88.41	86.70
Philosophy	77.41	70.65
College Computer Science	48.48	42.42
Average	59.34	56.37

Table 1. Evaluation results of different models on the subtasks of MMLU, using the script from chain-of-thought-hub [14].

4.3. Composibliity

As we have not modified the original LLM, our system can natively support the composition with question answering and image generation interleavedly.

5. Evaluation

Will prompting harm the inherent abilities of LLM? We provide a preliminary evaluation if the iT2I prompt harms the inherent abilities of LLM. As previously shown in Fig. 1, our prompting technique would not cause severe degradation in the LLM abilities. We can still ask LLMs for either question answering or code generation as before. To further investigate the impact of the iT2I prompt, we perform an ablation study on five subtasks of MMLU [19], comparing the models with and without the iT2I prompt. The results are provided in Tab. 1, it can be observed that the iT2I prompt only brings minor degradations.

Comparsion of different LLM. We evaluate our approach with different LLMs, including commerical ser-

Figure 7. Qualitative comparison of interactive text-to-image generation by prompting different LLM.

Interactive Image Generation

I want to draw a picture for my article, the theme is Spring Festival, do you have any idea?

Certainly! How about a vibrant illustration of a traditional Spring Festival scene? It could feature a bustling street filled with colorful decorations, lanterns, and people celebrating. There could be lion dancers, dragon parades, and families enjoying festive food together. The atmosphere should be joyful and lively, capturing the spirit of the Spring Festival.

Looks great, could you draw it?

Certainly! Here is an illustration of a traditional Spring Festival scene:

Can you imagine another scene of setting off firecrackers

Certainly! Here is an illustration of a scene where firecrackers are being set off during the Spring Festival:

T2I Model: Stable Diffusion XL

Storytelling

I have listened a beautiful song called "sleepyhead", could you help me draw illustration for it? In the following, I will give you the lyric part by part.

In the year of our lord 1239. There once lived a girl of a royal line. The ancient stories do recall. She was the fairest of them all.

In a castle made of stone. Every night she slept alone. Any noise that would raise the dead. Couldn't wake her sleepyhead

...

Could you summarize the story with interleaved images?

Certainly! Here's a summary of the story based on the lyrics:

In the year 1239, there lived a girl from a royal lineage who was known for her unmatched beauty. She resided in a castle made of stone and slept alone every night, undisturbed by any noise.
....

T2I Model: Stable Diffusion XL

Concept Prototyping

I want to design a picture of a spaceship, could you give me some suggestions?

Certainly! Here are a few suggestions for designing a picture of a spaceship:
1.Classic Spaceship: Create a sleek and futuristic spaceship with a metallic exterior, streamlined shape, and glowing lights.
2.Retro Spaceship: Design a spaceship inspired by retro sci-fi movies, with a bulbous shape, fins, and retro color schemes like silver and red.
3.Alien Spaceship: Let your imagination run wild and design an otherworldly spaceship with unconventional shapes, organic textures, and vibrant colors.

I like Retro Spaceship, could you imagine a picture of it?

Certainly! Here is an image of a Retro Spaceship for your inspiration:

What is the prompt for it?

The prompt for the image of the Retro Spaceship is "A Retro Spaceship with a bulbous shape, fins, and a silver and red color scheme."

I don't want the fins

Apologies for the misunderstanding. Here is an updated image of a Retro Spaceship without the fins:

T2I Model: Stable Diffusion XL

Interactive Logo Design

Can you help me to design a logo for a text-to-image app?

Of course! I can definitely help you design a logo for your text-to-image app. Could you please provide me with some more details about your app and any specific ideas or preferences you have for the logo design?

It is call mini-dalle3

Great! So you would like the logo for your text-to-image app called "Mini-DALLE3". Could you please provide me with some more information about the style or concept you have in mind for the logo? Do you have any specific colors or symbols that you would like to incorporate?

I would like it to be flatten

Understood! You would like the logo for your Mini-DALLE3 app to have a flat design. Based on that, I will create a logo that incorporates simplicity and minimalism. Here is a concept for the Mini-DALLE logo:
Please let me know if you have any specific changes or additions you would like to make to the concept.

I want it to use orange as main color

Certainly! Here is an updated concept for the Mini-DALLE logo with orange as the main color:
Please let me know if you have any further feedback or if there are any other specific elements you would like to include in the logo design.

T2I Model: DALLE 3

Figure 8. Examples use cases of interactive text-to-image generation.

Interactive Image Generation	Storytelling	Concept Prototyping	Interactive Logo Design
A vibrant illustration of a traditional Spring Festival scene. The street is filled with colorful decorations, lanterns, and people celebrating. Lion dancers and dragon parades add to the festive atmosphere, while families gather to enjoy delicious food together. The scene is filled with joy and excitement, capturing the spirit of the Spring Festival.	A beautiful girl with flowing hair, dressed in regal attire, stands tall in a grand castle. Her royal lineage is evident in her graceful demeanor. The ancient stories speak of her unmatched beauty, making her the fairest of them all.	A Retro Spaceship with a bulbous shape, fins, and a silver and red color scheme. It has a streamlined design and a nostalgic feel, reminiscent of classic sci-fi movies. The spaceship is adorned with retro-style decals and has large windows to provide a view of the cosmos. It exudes a sense of adventure and exploration as it cruises through the vastness of space.	A minimalist logo with the text "Mini DALLE 3" written in a clean and modern font. The letters are arranged in a straight line, with each letter having equal spacing. The color palette consists of vibrant shades of blue and green, representing creativity and innovation.
In this illustration, a group of people is gathered in an open area, setting off firecrackers to celebrate the Spring Festival. The sky is filled with colorful bursts of light and smoke as the firecrackers explode, creating a festive and energetic atmosphere. The people are wearing traditional clothing and have joyful expressions on their faces as they enjoy the tradition of setting off firecrackers to ward off evil spirits and bring good luck for the coming year.	Inside a majestic castle made of stone, the girl sleeps peacefully in her grand bed. Despite any noise that could wake the dead, she remains undisturbed in her slumber, earning her the nickname "Sleepyhead."	A Retro Spaceship with a bulbous shape and a silver and red color scheme. It has a streamlined design reminiscent of classic sci-fi movies. The spaceship features a rounded cockpit with large windows, allowing the pilot to have a panoramic view of space. The exterior is adorned with retro-style decals and has a polished metallic finish. It exudes a sense of nostalgia and adventure as it travels through the cosmos.	A minimalist logo with the text "Mini DALLE 3" written in a clean and modern font. The letters are arranged in the same line, with each letter having equal spacing. The color palette consists of vibrant shades of green and orange representing creativity and innovation. The logo is placed on a white background, creating a clean and sleek look.

Table 2. Generated captions of the examples in Figure 8.

vices OpenAI GPT3.5 [7], GPT4 [36], Claude[3], and open-source LLAMA2-13B-Chat [48], Baichuan2-13B Chat [56], ChatGLM2-6B-Chat [10], Qwen-14B-Chat [3], InternLM-20B-Chat [46]. As shown in Fig. 7, all commercial LLMs successfully generate the images with appropriate corrsponding text (interleaved) responses. This indicates that our prompting approach could be a simple yet effective way to rapidly augment existing LLMs with iT2I ability. Nevertheless, the results are less satisfactory for the open-source LLMs. Overall, Baichuan2 [56] generates the best results, while Qwen and InternLM tend to refuse to generate images even if they are prompted to do so. ChatGLM2 could generate an image but the correspondence is incorrect.

iT2I Examples. Here, we show a number of iT2I examples, which cover different use scenarios from single-turn/multi-turn image generation to interleaved text-image storytelling. The results are shown in Fig. 8 and Tab. 2.

6. Conclusion

In conclusion, this paper introduces the concept of interactive text-to-image (iT2I) and presents an approach to augmenting existing large language models for this task. Our evaluation shows that this approach enables convenient iT2I capabilities without significant degradation of the models' inherent capabilities. This work has the potential to enhance user experiences in human-machine interactions and elevate the image quality of next-generation T2I models, offering promising directions for future research and development.

[3] https://claude.ai

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. *arXiv preprint arXiv:2309.06933*, 2023. 3

[2] AUTOMATIC1111. Stable Diffusion Web UI, 2022. 2

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. 9

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 3

[5] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models, 2023. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3, 5

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3, 9

[8] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023. 6

[9] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 3, 5

[10] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 9

[11] Yuki Endo. User-controllable latent transformer for stylegan image layout editing. *Computer Graphics Forum*, 41(7): 395–406, 2022. 3

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[13] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023. 2, 3

[14] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023. 6

[15] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[16] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 3, 5

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[24] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023. 3, 5

[25] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 3

[26] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023. 6

[27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[28] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. *arXiv preprint arXiv:2309.07906*, 2023. 3

[29] Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, and Yi Jin. Freedrag: Point tracking is not you need for interactive point-based image editing. *arXiv preprint arXiv:2307.04684*, 2023. 3

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[31] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3

[32] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 6

[33] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 3

[34] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[35] OpenAI. Dall·e 3 system card. 2023. 2

[36] OpenAI. Gpt-4 technical report, 2023. 3, 5, 9

[37] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 6

[43] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[44] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 3

[45] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 3

[46] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023. 3, 9

[47] Haileleol Tibebu, Aadin Malik, and Varuna De Silva. Text to image synthesis using stacked conditional variational autoencoders and conditional generative adversarial networks. In *Science and Information Conference*, pages 560–580. Springer, 2022. 3

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3, 5, 9

[49] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023. 3

[50] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 5

[51] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 3

[52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3

[53] Jingxuan Wei, Shiyu Wu, Xin Jiang, and Yequan Wang. Dialogpaint: A dialog-based image editing model. *arXiv preprint arXiv:2303.10073*, 2023. 3

[54] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2023. 3, 5

[55] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 2, 3

[56] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 3, 9

[57] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3, 6

[58] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 3

[59] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[60] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 6

[61] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 3

[62] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[63] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2023. 5

[64] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

[65] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 5