LLMGA: Multimodal Large Language Model based Generation Assistant

Bin Xia[1], Shiyin Wang[2], Yingfan Tao[2], Yitong Wang[2], and Jiaya Jia[1]

[1] The Chinese University of Hong Kong
[2] ByteDance Inc
https://llmga.github.io/

Abstract. In this paper, we introduce a Multimodal Large Language Model-based Generation Assistant (LLMGA), leveraging the vast reservoir of knowledge and proficiency in reasoning, comprehension, and response inherent in Large Language Models (LLMs) to assist users in image generation and editing. Diverging from existing approaches where Multimodal Large Language Models (MLLMs) generate fixed-size embeddings to control Stable Diffusion (SD), our LLMGA provides a detailed language generation prompt for precise control over SD. This not only augments LLM context understanding but also reduces noise in generation prompts, yields images with more intricate and precise content, and elevates the interpretability of the network. To this end, we curate a comprehensive dataset comprising prompt refinement, similar image generation, inpainting & outpainting, and instruction-based editing. Moreover, we propose a two-stage training scheme. In the first stage, we train the MLLM to grasp the properties of image generation and editing, enabling it to generate detailed prompts. In the second stage, we optimize SD to align with the MLLM's generation prompts. Additionally, we propose a reference-based restoration network to alleviate texture, brightness, and contrast disparities between generated and preserved regions during inpainting and outpainting. Extensive results show that LLMGA has promising generation and editing capabilities and can enable more flexible and expansive applications in an interactive manner.

Keywords: Interactive generation and editing · LLM · Diffusion Model

1 Introduction

Artificial Intelligence Generated Content (AIGC) has witnessed remarkable advancements, particularly propelled by the evolution of large language models (LLMs) [8, 10, 54] for text generation and diffusion models (DMs) [21, 42] for image generation. LLMs, in particular, have received considerable acclaim for their exceptional ability to comprehend, reason, make decisions, possess extensive knowledge, and generate text with unparalleled accuracy and fluency.

Recent studies have begun delving deeper into Multimodal Large Language Models (MLLMs) [1, 69] built upon LLMs, aiming to empower LLMs to comprehend inputs extending beyond text. For example, BLIP-2 [30] and LLaVA [33]

Fig. 1: Some examples of LLMGA for assisting in image generation and editing. **(1)** T2I generation. LLMGA can refine the user's generation prompt to produce more vivid and vibrant images. **(2)** Similar image generation. LLMGA can understand the component and layout of the input images and generate a similar image. **(3)** Inpainting & Outpainting. LLMGA can provide detailed generation prompts based on user preferences and input images. **(4)** Instruction based editing. LLMGA can understand user instructions and realize accurate editing. **(5)** Interactive image generation and editing exemplify the comprehensive capabilities of LLMGA. Users can design satisfactory images by engaging in interactions with LLMGA, leveraging its vast knowledge.

employ visual encoders to transform images into input embeddings, enabling them to be used as prompts alongside text input for the LLM, thus achieving LLMs with the visual modality. Furthermore, recent works focused on extending the capabilities of LLMs to generate multimodal outputs. For example, GILL [29] involves instructing LLMs to predict fixed-size visual embeddings aligned with CLIP [40] space to control the Stable Diffusion [42] (SD) for image generation.

However, existing works [29, 52] merely focus on enabling LLM to output images but do not aim to assist users in generating or editing images to enhance quality. In this paper, we aim to develop a Multimodal Large Language Model-based Generation Assistant (LLMGA) to better assist image generation models, making them more user-friendly and capable of producing high-quality images. In contrast to certain methods [29, 52] that leverage MLLMs to predict fixed-size visual embeddings for implicit SD control, our approach is straightforward. We guide the generation of SD using detailed language prompts from MLLM based on five reasons. **(1)** The embeddings predicted by the MLLM are often filled with noise. This can be filtered out by mapping them to a fixed language domain, enabling precise control of SD. **(2)** Detailed language prompts can make the network more transparent and interactive, allowing users to understand MLLM's thoughts for generating images. **(3)** MLLM is pre-trained on vast textual datasets. Explicit language prompts rather than implicit embeddings are more advantageous for MLLM to generate prompts and comprehend context. **(4)**

Dynamic-sized language prompt facilitates the addition of generation requests during interactions. **(5)** Training is more simple and efficient.

However, we face several challenges: **(1)** MLLM may reject the execution of generation instructions due to its nature as a language assistant. **(2)** MLLM lacks a comprehensive understanding of image generation and editing, and cannot provide an accurate and detailed generation prompt. **(3)** Determining which part of texts generated by MLLM to guide SD generation. **(4)** SD's CLIP encode only 75 tokens. Additionally, SD is trained on short captions, whereas our LLMGA typically generates detailed prompts exceeding 150 tokens. This discrepancy poses a challenge for SD in following the detailed prompt of LLMGA.

To this end, we have devised a two-stage training scheme. First, we construct a training dataset: prompt refinement, similar image generation, inpainting & outpainting, and visual question answering. We then train LLMGA on these four datasets to cultivate four fundamental capabilities: **(1)** For concise user prompts, LLMGA can refine the generation of intricate details, encompassing attire, background, and characters. **(2)** LLMGA can precisely regenerate an image it observes. **(3)** LLMGA can generate or refine prompts for inpainting & outpainting based on its understanding of the image. **(4)** LLMGA can generate accurate prompts for instruction-based editing according to users' requirements and given images. Additionally, we make LLMGA use special symbols `<gen_img>` and `</gen_img>` to distinguish generation prompts and responses. In the second stage, we freeze the parameters of LLMGA's MLLM and initiate joint training with the SD. This process enables the SD to acclimate to the detailed prompt produced by the MLLM. Notably, when the input token count exceeds 75, we iteratively apply the CLIP [40] encoder to the surplus tokens.

Moreover, we have identified noticeable disparities in texture, contrast, and brightness between the newly generated and preserved sections in inpainting & outpainting. Therefore, we propose a Diffusion-based Reference Restoration Network (DiffRIR). Specifically, aside from images generated by SD, we add masked images as reference inputs into DiffRIR. This enables the DiffRIR to refer to the texture, contrast, and brightness of the preserved regions for restoration. Additionally, we introduce perturbations to contrast and brightness during training, enabling DiffRIR to correct contrast and brightness disparities in the images.

As shown in Fig. 1, LLMGA is a unified and interactive framework for image generation and editing, endowed with a wide array of capabilities: **(1)** LLMGA leverages its extensive world knowledge and powerful reasoning abilities to assist image generation and editing and significantly improve results. **(2)** LLMGA can be integrated with external plugins, like ControlNet. **(3)** Most importantly, users can interact with LLMGA to design satisfying images in a more convenient, flexible, and enjoyable way. In summary, our contributions are as follows:

- We proposed LLMGA, a simple yet powerful interactive generation and editing framework. Experiments affirm the efficacy of LLMGA in enhancing generation and editing thanks to its vast knowledge and interactive features. Plus, LLMGA can integrate with external plugins for wider applications.

- We construct a training dataset, including four parts: prompt refinement, similar image generation, inpainting & outpainting, and instruction-based editing. This enhances LLMGA's comprehension of generation and editing tasks while standardizing response formats.
- We proposed a restoration network DiffRIR, which introduces reference images and training perturbations to contrast and brightness. DiffRIR can alleviate texture, contrast, and brightness discrepancies between newly generated and preserved regions for edited images.
- Open-source. The following assets are released: the generated data, the codebase for model training, the model checkpoint, and a demo.

2 Related Work

Diffusion Model. Diffusion Models (DMs) [2,4,7,13,19,21,22,28,35,38,49,51] have achieved remarkable results in image generation. DMs adopt a parameterized Markov chain to optimize the lower variational bound on the likelihood function. In this way, it can generate realistic images from Gaussian noise. After that, several DM methods [3,5,6,12,16,17,26,27,34,39,44,45,55,59,60,62] have been tailored to enhance the text-to-image (T2I) generation and editing. Notably, GLIDE [36] pioneered the incorporation of text features into transformer blocks during the denoising process. Subsequently, DALL-E [41], Imagen [46], and Stable Diffusion [42] have made substantial strides in improving T2I generation. Subsequently, some works [43, 66, 67] introduced conditioning controls to the DMs to facilitate a more convenient and precise manipulation of the generation process. Overall, enhancing the user-friendliness of DMs is a key focus within the community. In this paper, we introduce LLMGA, leveraging the extensive knowledge and powerful reasoning capabilities of LLM to facilitate users in achieving more easily attainable and satisfactory image designs.

Multimodal Large Language Models. Recently, LLMs have undeniably made profound impacts and revolutions within the entire AI community and beyond. For example, exemplary LLMs, such as ChatGPT and GPT4 [37], have showcased remarkable abilities in comprehension, reasoning, responses, and knowledge reservoirs. Subsequently, a range of LLMs [11], including Vicuna [9], LLaMA [54], and Alpaca [53] have been released as open-source models, substantially propelling advancements of the community.

Afterward, the community began focusing on the development of the Multimodal Large Language Model [14, 15, 18, 23, 63, 64, 69]. They aim to enable LLMs to comprehend both images and text and provide textual responses. For instance, Flamingo [1] encodes images and feeds them into the LLM's attention layer. BLIP-2 [30] employs Q-Former to encode input images into queries. Additionally, LLaVA [33] leverages CLIP [40] to encode images into visual embeddings, which are then concatenated with text embeddings.

Recent concurrent works, such as Next-GPT [58], have extended the capabilities to encompass audio and video modalities. Moreover, Visual-ChatGPT [57] and HuggingGPT [48] make LLMs act as agents capable of invoking various pretrained visual models to achieve MLLM. However, these works focus on making

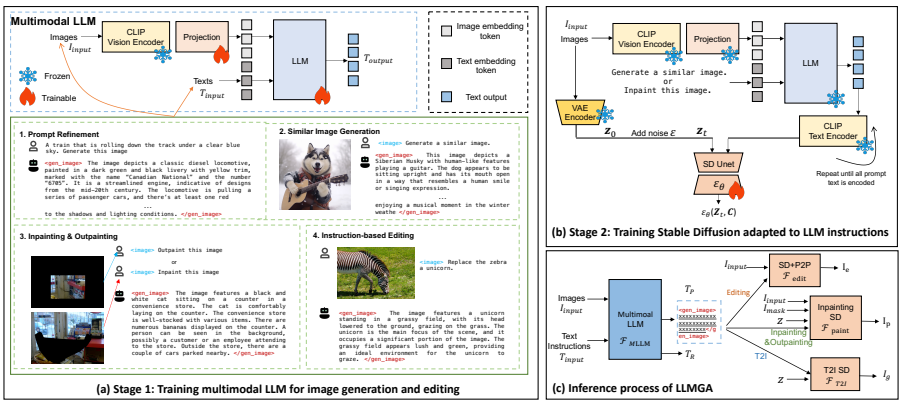

Fig. 2: The overview of LLMGA. **(a)** In the first training stage, we train the MLLM to produce generation prompts based on provided instructions. Moreover, we construct a training dataset including four categories: prompt refinement, similar image generation, inpainting & outpainting, and instruction-based editing. **(b)** In the second training stage, we optimize SD to adapt to the detailed generation prompts from MLLM. **(c)** In the inference stage, LLMGA can respond to user queries and assist in various tasks, such as image generation, inpainting & outpainting, and editing.

LLM determine the combined invocation of modules (such as detection, recognition, and generation) to fulfill user requirements. However, these methods are not tailored for generation and editing and use concise prompts that lack the capability to enhance results. Thus, we propose LLMGA, which is designed to assist with various image generation and editing tasks. It can achieve satisfactory results by strong reasoning capability and flexible interaction with users.

3 Methodology

3.1 Overview of LLMGA

In this paper, we aim to design a MLLM-based Generation Assistant (LLMGA). Our LLMGA produces detailed language-based generation prompts to control SD rather than predicting fixed-sized visual embeddings [29, 52] to govern SD. This has five advantages: **(1)** Visual embeddings contain noise, and mapping them to the language domain can filter out this noise, enabling precise SD control. **(2)** Language-based generation prompts facilitate users in understanding the LLMGA's thoughts, enhancing interaction. **(3)** Dynamic-sized language-based generation prompt enables the addition of generation requests. **(4)** MLLM is pre-trained on textual datasets. Language prompts rather than implicit visual embeddings are more advantageous for MLLM to generate accurate prompts and comprehend context. **(5)** Training is simpler and more efficient.

However, we need to address several issues: **(1)** As a language assistant, MLLM may decline the execution of generation instructions. **(2)** MLLM lacks

a nuanced understanding of image generation and editing, and cannot produce precise and detailed generation prompts. **(3)** MLLM needs to decide which part of the output text serves as generation prompts to guide generation. **(4)** SD's CLIP encode only 75 tokens. Moreover, SD is primarily trained on short captions, while detailed prompts generated by LLMGA may exceed 150 tokens. This disparity makes it hard for SD to understand the instructions from LLMGA. To address the aforementioned challenges, we construct a training dataset and design two-stage training schemes to train the MLLM (Sec. 3.2) and SD (Sec. 3.3).

The network structure and pipeline of LLMGA are illustrated in Fig. 2. Specifically, as shown in Fig. 2 (a) and (c), the images $\mathbf{I}_{input}$ are encoded into image embeddings by CLIP vision encoder and a projection layer. Subsequently, the image embedding is concatenated with the text embedding and fed into the LLM to obtain text output $\mathbf{T}_{output}$. This process can be formulated as:

$$\mathbf{T}_{output} = \mathcal{F}_{MLLM}(\mathbf{T}_{input}, \mathbf{I}_{input}), \tag{1}$$

where $\mathbf{T}_{input}$ indicates the input text instructions from users. It is notable that $\mathcal{F}_{MLLM}$ can process only $\mathbf{T}_{input}$ as input.

The text output $\mathbf{T}_{output}$ can comprise two components: text response $\mathbf{T}_R$ and generation prompt $\mathbf{T}_P$. To distinguish between $\mathbf{T}_R$ and $\mathbf{T}_P$, we adopt new special tokens, *i.e.*, `<gen_img>` and `</gen_img>`, to encompass $\mathbf{T}_P$.

We present $\mathbf{T}_R$ as the immediate text response to users. Concurrently, $\mathbf{T}_P$ is further fed into the subsequent SD to guide T2I generation (Eq. 2), inpainting & outpainting (Eq. 3), and instruction-based image editing (Eq. 4).

$$\mathbf{I}_g = \mathcal{F}_{T2I}(\mathbf{T}_P, \mathbf{Z}), \tag{2}$$

$$\mathbf{I}_p = \mathcal{F}_{Paint}(\mathbf{I}_{input}, \mathbf{I}_{mask}, \mathbf{T}_P, \mathbf{Z}), \tag{3}$$

$$\mathbf{I}_e = \mathcal{F}_{Edit}(\mathbf{I}_{input}, \mathbf{T}_P), \tag{4}$$

Where $\mathbf{Z}$ denotes the random Gaussian noise. Furthermore, to ensure the encoding of all $\mathbf{T}_P$ for SD, we iteratively run the CLIP text encoder until all prompts are encoded. For inpainting & outpainting, except the input image $\mathbf{I}_{input}$, an additional mask $\mathbf{I}_{mask}$ are essential inputs for the inpainting SD to specify the region requiring generation. For instruction-based image editing, $\mathcal{F}_{Edit}$ performs inversion [25, 50] and prompt-to-prompt [20] based on T2I SD.

3.2 MLLM Training

As described in Sec. 3.1, original MLLMs are specifically designed and trained as language assistants, but they lack the proficiency to assist in image generation, and editing. Notably, considering the input of images and the performance of open-sourced MLLMs, employing few-shot learning to guide the model to achieve the desired results proves challenging and inefficient. Therefore, it is crucial to train the MLLM to serve as proficient assistants in image generation and editing

tasks, understanding the expected response formats and enhancing their comprehension of image generation and editing properties. To this end, as depicted in Fig. 2 (a), we construct a training dataset consisting of four categories:

(1) *Prompt Refinement.* We establish this dataset to cultivate the prompt refinement ability of the MLLM. Specifically, we utilize GPT4-V to furnish detailed descriptions of images in MSCOCO [31]. These detailed descriptions, along with the original MSCOCO brief descriptions, constitute a training text pair. During training, we input the brief MSCOCO captions and randomly select and append a generation instruction. When a generation instruction is included in the prompt, we add `<gen_img>` and `</gen_img>` on the later description.

(2) *Similar Image Generation.* We select images from the MSCOCO dataset along with corresponding detailed descriptions generated by GPT4-V to create the Similar Image Generation dataset. During training, we input the images along with a generation instruction. In cases where a generation instruction is provided, we add `<gen_img>` and `</gen_img>` on the subsequent description.

(3) *Inpainting & Outpainting.* We use pairs of detailed descriptions and images from the Similar Image Generation dataset. During training, we input masked images with inpainting or outpainting instructions. Besides, we include `<gen_img>` and `</gen_img>` on the subsequent description.

(4) *Instruction-based editing.* We fine-tune Mixtral-8x7B [24] to enable it to generate editing data based on detailed descriptions from MSCOCO. Subsequently, we clean the generated data. During training, we can input images or original captions and provide corresponding editing instructions, aiming to train LLMGA to output the target caption. Additionally, we include `<gen_img>` and `</gen_img>` tags on the target caption during training.

During training, alongside the above data, we integrated the image designing and Alpaca [53] dataset, to enhance LLMGA's question-answering (QA). Notably, we excluded certain visual multimodal incompatible question-answer pairs from it. Furthermore, we incorporated the LLaVA v1.5 mix665k dataset [32] to endow LLMGA with Visual Question Answering (VQA) capabilities. We provide more details in the supplementary material.

As illustrated in Fig. 2 (a), we freeze the CLIP vision encoder and optimize the projection layer and LLM. The model is trained end-to-end using the auto-regressive cross-entropy loss ($\mathcal{L}_{MLLM}$) for text generation. Given the ground-truth targets $\mathbf{T}_{GT}$, this loss can be formulated as:

$$\mathcal{L}_{MLLM} = \mathbf{CE}\left(\mathbf{T}_{output}, \mathbf{T}_{GT}\right). \tag{5}$$

3.3 Stable Diffusion Training

As described in Sec. 3.1, the original SD's CLIP text encoder only encodes 75 tokens, which cannot handle the entire MLLM's generation prompt. Moreover, the original SD is trained on brief captions, which cannot fully understand the generation prompts. Thus, we repeatedly use the CLIP text encoder to encode all instances of $\mathbf{T}_P$ for SD. Besides, we train the T2I SD model and the inpainting SD model, respectively. For both generation and inpainting & outpainting tasks,

the generation prompts of MLLM are detailed descriptions of images. Therefore, during training, we can instruct MLLM to provide a detailed description $\mathbf{T}_P$ for images from the LAION-Aesthetics [47] and MSCOCO datasets. Subsequently, $\mathbf{T}_P$ is fed into T2I SD or inpainting SD for joint training. Notably, we only optimize the SD unet while freezing the parameters of other networks. To accelerate the training, we record the prompt $\mathbf{T}_P$ of MLLM to avoid redundant calculations. The model is trained using SD loss (Eq. 6). For instruction-based editing, we directly adopt the pre-trained T2I SD model.

$$\mathcal{L}_{SD} = \mathbb{E}_{\mathbf{Z}_t,\mathbf{C},\epsilon,t}\left(\|\epsilon - \epsilon_\theta\left(\mathbf{Z}_t, \mathbf{C}\right)\|_2^2\right), \tag{6}$$

where $\mathbf{Z}_t = \sqrt{\overline{\alpha_t}}\mathbf{Z}_0 + \sqrt{1-\overline{\alpha_t}}\epsilon$ represents the noised feature map at timestep t. Ground truth images are encoded into latent space to derive $\mathbf{Z}_0$. Here, $\epsilon \in \mathcal{N}(0, \mathbf{I})$ represents Gaussian noise, and ϵ_θ refers to the SD unet. $\mathbf{C}$ indicates the conditional information. For T2I generation, $\mathbf{C}$ is $\mathbf{T}_P$. For inpainting and outpainting, $\mathbf{C}$ contains $\mathbf{T}_P$, the mask, and the VAE-encoded masked image.

3.4 Restoration Network Training

For SD inpainting & outpainting, we observed noticeable disparities between the preserved and newly generated regions in the edited images. To enhance the consistency between the newly generated and the preserved regions, we introduce a reference-based restoration scheme. Specifically, existing restoration methods [56, 60, 61] take low-quality (LQ) images as input and produce high-quality (HQ) images, but they often do not leverage the preserved information from the given masked image. Different from them, we concatenate the LQ image I_{LQ} and the masked image, *i.e.*, $(1 - \mathbf{I}_{mask})\mathbf{I}_{GT}$, as inputs and feed them into the restoration model $\mathcal{F}_R$. This process can be formulated as:

$$\mathbf{I}_{HQ} = \mathcal{F}_R(concat(\mathbf{I}_{LQ}, (1 - \mathbf{I}_{mask})\mathbf{I}_{GT})). \tag{7}$$

To further mitigate the brightness and contrast disparities, we introduced additional color degradation (*i.e.*, random brightness and contrast disturbance) into the training process of the restoration model, which is formulated as:

$$\mathcal{D}_2(\mathbf{x}) = c_1\mathbf{x} + c_2, \tag{8}$$

$$\mathbf{I}_{LQ} = \mathcal{D}_2(\mathcal{D}_1(\mathbf{I}_{GT})), \tag{9}$$

where $\mathcal{D}_2$ indicates contrast and brightness disturbance, $\mathbf{x}$ denotes the input image. c_1 is the contrast gain, randomly varied within the range of $[0.94, 1.06]$, while c_2 is the brightness bias, randomly varied within the range of $[-0.05, 0.05]$. $\mathcal{D}_1$ represents the real-world degradation process used in restoration model [56, 60, 61]. $\mathbf{I}_{GT}$ is ground truth image, and $\mathbf{I}_{LQ}$ is LQ image. Here, we adopt the network structure of the SOTA restoration model DiffIR [60] and apply our schemes to it to obtain our Diffusion-based Reference Restoration Network (DiffRIR).

4 Experiments

4.1 Implementation Details

For the first stage of training, we train the MLLM (including projection and LLM) on 8×A100. The batch size is set to 128. Besides, the training datasets include VQA (LLaVA v1.5 mix665k), QA, prompt refinement, similar image generation, inpainting & outpainting, and instruction-based editing.

For the second stage of training, we adopt the Stable Diffusion 1.5 (SD1.5) as the initial image generation or inpainting & outpainting model. We train these models on 8×A100. The batch size is set to 32.

For the restoration network, we train DiffRIR on 8×V100. The batch size is set to 64. Please check more details in the supplementary material.

4.2 Experimental Results

Table 1: Quantitative comparison on **T2I** generation on the MSCOCO [31] dataset.

Method	FID↓	IS↑
SD1.5	24.0081	35.99
GILL [29]	25.1123	34.20
LLMGA-7b(SD1.5)	23.5946	37.12
LLMGA-7b(SD1.5-ft, Ours)	<u>18.5234</u>	<u>41.04</u>
LLMGA-13b(SD1.5)	23.5828	37.58
LLMGA-13b(SD1.5-ft, Ours)	**18.4063**	**41.16**

Evaluation on T2I Generation. The results are shown in Tab. 1. Notably, SD1.5 is the original Stable Diffusion 1.5 while SD1.5-ft indicates the finetuned SD1.5 in LLMGA. We also compare our LLMGA with the recently proposed multimodal generative model GILL [29]. **(1)** Comparing SD1.5, our LLMGA-7B achieves notable 5.4847 FID and 5.05 IS improvements, underscoring the effectiveness of LLMGA. **(2)** Moreover, our LLMGA-7B significantly outperforms GILL. **(3)** In the 3rd and 4th rows of Tab. 1, we use LLMGA-7B to refine the short MSCOCO caption to the detailed generation prompt and send it to the SD1.5 and SD1.5-ft, respectively. Our LLMGA-7B with SD1.5-ft achieves significant 5.0712 FID and 3.92 IS improvements, respectively. This demonstrates our second-stage training can help SD1.5-ft to better follow detailed prompts from MLLM. **(4)** Comparing the 4th and 6th rows of Tab. 1, LLMGA-13B exhibits better performance than LLMGA-7B due to its superior reasoning ability.

The qualitative results are shown in Fig. 3. **(1)** LLMGA excels in refining prompts by incorporating details to generate visually rich and pleasing images. For example, in the first row, LLMGA crafts a battle attire for the husky, depicting engaging scenarios of battling monsters. This makes user usage more convenient, eliminating the need for them to think about generating image details themselves. **(2)** LLMGA can leverage its extensive knowledge base to generate images, even for concepts users may not be familiar with, like a space elevator.

Evaluation on Instruction-based Editing. For instruction-based editing, we utilize LLMGA to provide a detailed description of the edited image based on the input image and user editing instructions. We then employ Direct Inversion [25]

Fig. 3: T2I visual comparison. LLMGA can produce accurate and high-quality results.

Table 2: Quantitative comparison for **image editing** on MagicBrush test set [65].

Methods	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
InstructPix2Pix	0.1197	0.0416	0.8442	0.7252	0.2909
MagicBrush	**0.0647**	0.0224	**0.9293**	**0.8913**	0.2979
LLMGA (Ours)	0.0814	**0.0218**	0.8936	0.8768	**0.3137**

and prompt-to-prompt [20] methods to obtain the edited image. **(1)** The quantitative results are presented in Tab. 2. Notably, our LLMGA did not train SD for image editing like InstructPix2Pix [7] and MagicBrush [65]. However, our zero-shot performance on the MagicBrush test set surpassed that of InstructPix2Pix, achieving performance similar to that of MagicBrush. **(2)** Additionally, our LLMGA offers a superior user experience in interactive editing, allowing image modifications to be carried out conversationally. In contrast, InstructPix2Pix only supports input via instructions and output as images. Visualized results are shown in Fig. 4, by leveraging the powerful reasoning capabilities of LLM, our LLMGA can provide more accurate and reasonable editing results.

Evaluation on Inpainting and Outpainting. The results are shown in Tab. 3. **(1)** Comparing the 1st and 3rd rows of Tab. 3, our LLMGA-7B achieves significant improvements of 2.5681 FID over the SD1.5 in outpainting under wide masks. **(2)** In the 2nd and 3rd rows of Tab. 3, we make LLMGA imagine the complete generation prompts for the given masked images, which are then input into the later SD. Notably, our LLMGA-7B (with SD1.5-ft) demonstrates a significant FID improvement over LLMGA-7B (with SD1.5) in both outpainting and inpainting. This demonstrates the second stage of training makes SD better follow the prompts from MLLM. **(3)** Comparing the 3rd and 5th rows, our LLMGA-13B outperforms LLMGA-7B due to its superior reasoning capabilities.

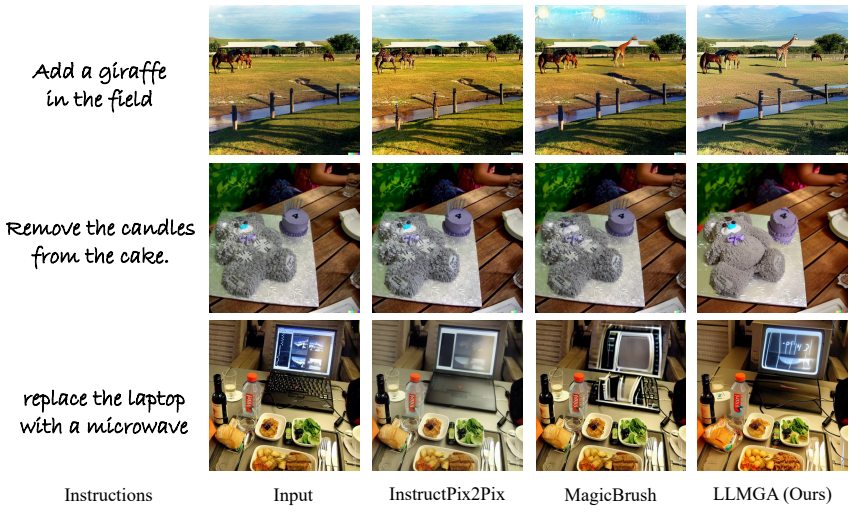

Fig. 4: Visual comparison on **instruction-based editing**.

Table 3: Quantitative comparison for **outpainting & inpainting** on Places [68].

Method	Outpainting				Inpainting			
	Narrow Masks		Wide Masks		Narrow Masks		Wide Masks	
	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓
SD1.5	2.0167	0.2283	5.0090	0.3734	1.0795	0.1236	1.2855	0.1434
LLMGA-7b (SD1.5)	1.5530	0.2263	3.2630	0.3688	1.0692	0.1233	1.0983	0.1427
LLMGA-7b (SD1.5-ft, Ours)	1.2973	0.2215	2.4409	0.3616	0.8027	0.1171	0.9807	0.1405
LLMGA-13b (SD1.5)	1.5631	0.2263	3.0845	0.3679	1.0326	0.1228	1.0978	0.1426
LLMGA-13b (SD1.5-ft, Ours)	**1.2160**	**0.2210**	**2.3663**	**0.3609**	**0.7992**	**0.1166**	**0.9780**	**0.1400**

The qualitative results are shown in Fig. 5. We can see that LLMGA can deduce and imagine complete images based on masked input images. For example, in the 3rd row of Fig. 5, LLMGA can infer the presence of wind turbines on the mountain based on the given environment. Overall, LLMGA's powerful reasoning capability and extensive knowledge can assist users in conveniently making accurate and visually pleasing inpainting and outpainting.

Evaluation on ControlNet. Our LLMGA demonstrates exceptional scalability, enabling integration with external plugins like ControlNet [66]. Here, we utilize LLMGA to create detailed prompts derived from input images and user requirements, working alongside ControlNet to guide image generation. As depicted in Fig. 6, our LLMGA significantly enhances the diversity and richness of outcomes in picture reference-guided image generation. Furthermore, more external plugins can also be integrated into the interactive framework of LLMGA,

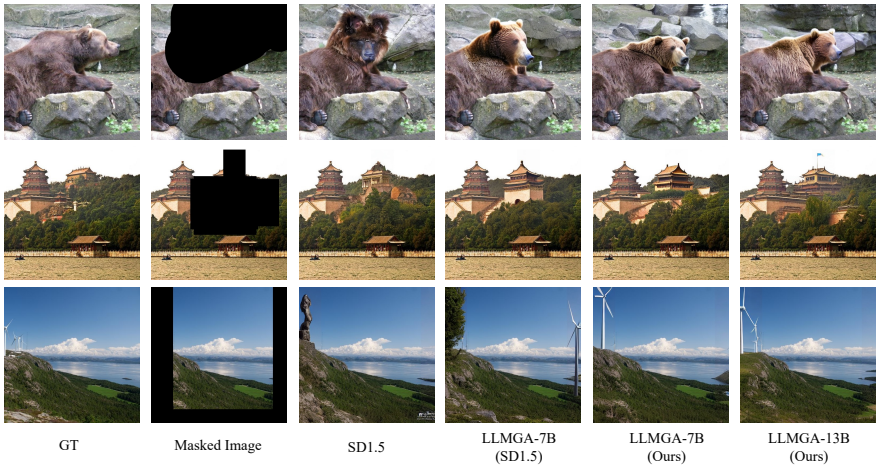

Fig. 5: Visual comparison on **inpainting and outpainting**.

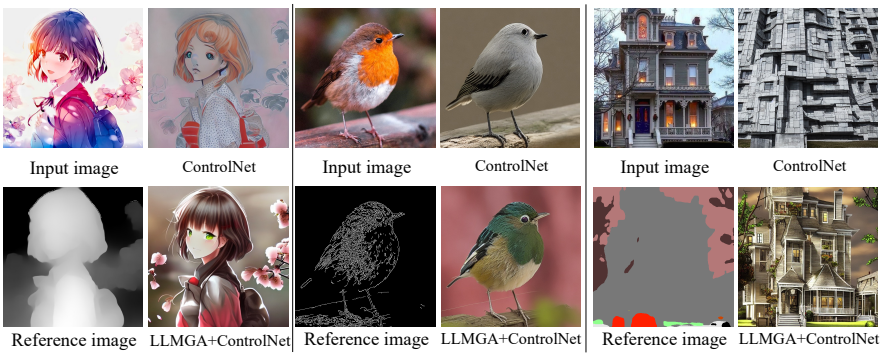

Fig. 6: Visualization of LLMGA plus ControlNet. Our LLMGA can enhance the details in generated images, producing visually pleasing images.

combining with LLMGA's reasoning design capabilities and all previous functions to allow for a broader range of creative and engaging applications.

Evaluation on Image Restoration. The results are shown in Tab. 4. For comparisons, we validate DiffIR [60] and our DiffRIR on the outpainting image generated by LLMGA. **(1)** Comparing the 2nd and 3rd rows of Tab. 4, it is evident that introducing a reference scheme can significantly improve restoration performance. **(2)** When comparing the 3rd and 4th rows of Tab. 4, it can be observed that introducing color degradation helps alleviate the bright and contrast distortion caused by SD. **(3)** Comparing the 1st and 4th rows of Tab. 4, our DiffRIR yields significant improvement, validating the effectiveness of DiffRIR.

As shown in Fig. 7, our DiffRIR (*i.e.*, DiffRIR_2 in Tab. 4) can alleviate the texture, brightness, and contrast discrepancies, and generate realistic details.

Table 4: Quantitative comparison on **image restoration** for Places [68] outpainting.

Method	Reference	Color Degradation	FID↓	LPIPS↓
LLMGA-7B	✗	✗	2.4409	0.3616
LLMGA+DiffIR [60]	✗	✗	2.3587	0.3612
LLMGA+DiffRIR$_1$	✓	✗	2.2993	0.3609
LLMGA+DiffRIR$_2$ (Ours)	✓	✓	2.2687	0.3607

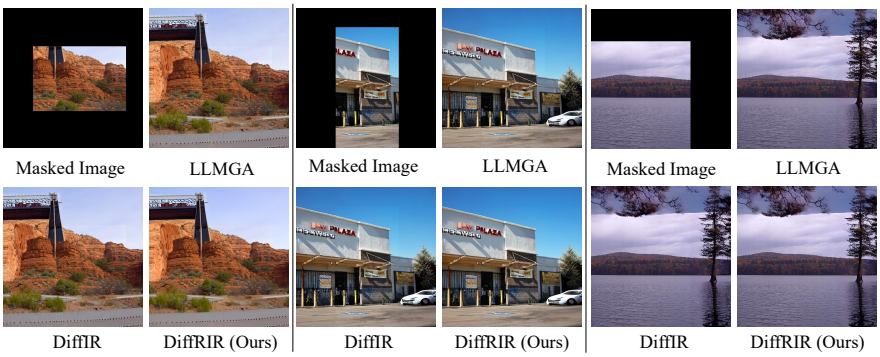

Fig. 7: Visual comparison of image **restoration** methods. DiffRIR can alleviate the texture, contrast, and brightness disparities in inpainting & outpainting results.

Control SD using detailed language prompt or embedding? The results are shown in Fig. 5. We compare two approaches: GILL [29], which makes LLM estimate a fix-sized embedding to control SD generation, and LLMGA Embedding, a variant of LLMGA where the language prompt is replaced with embedding, undergoing the same training process as LLMGA. The evaluation is conducted on MSCOCO by instructing these methods to generate images with the same prompts in multiple times in conversation form. **(1)** The quality of generated images (Fig. 5) in embedding-based methods (*i.e.*, GILL and LLMGA Embedding) deteriorate rapidly as the number of conversation turns increases. In contrast, our LLMGA remains unaffected. This discrepancy arises from the inherent noise present in the embeddings predicted by LLM. As the number of conversation turns rises, these generated embeddings

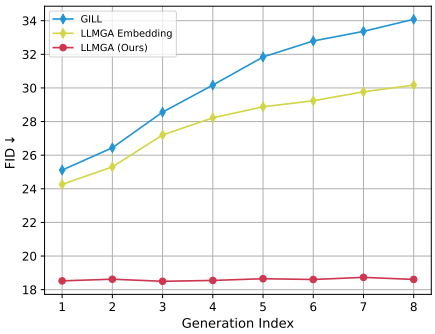

Table 5: T2I performance comparison of SD control schemes based on detailed language prompt and embedding.

Table 6: Datasets comparison. We conducted comparisons of FID on T2I and outpainting. The ✓signifies the utilization of the complete dataset during training. Conversely, the absence of the ✓indicates a reduction to only 10% of the original datasets.

Method	Prompt Refinement	Inpainting &Outpainting	Similar Image Generation	Instruction-based Editing	T2I	Outpainting
$LLMGA_1$		✓	✓	✓	21.4460	2.6533
$LLMGA_2$	✓		✓	✓	19.1914	3.1054
$LLMGA_3$	✓	✓		✓	19.5698	2.6612
$LLMGA_4$	✓	✓	✓		19.0642	2.8047
$LLMGA_5$ (Ours)	✓	✓	✓	✓	18.5234	2.4409

integrate with the preceding conversations, introducing even more noise. This poses challenges for the precise control of SD-generated content. Our LLMGA addresses this issue by mapping the embedding to the fixed language domain, effectively eliminating such noise. **(2)** Additionally, LLMGA Embedding also outperforms GILL, indicating that the prompt size used to guide SD generation should be adaptive in content, rather than a fixed size.

Contribution of Training Data. To assess the impact of training data, we downsized one of the four training datasets in $LLMGA_5$ to 10% of its original magnitude, ensuring that LLMGA remains capable of furnishing responses in the prescribed format. The results are shown in Tab. 6. It is evident that prompt refinement, inpainting & outpainting, and instruction-based editing datasets enhance LLMGA's comprehension of image generation and editing properties, resulting in superior images. Moreover, comparing $LLMGA_3$ and $LLMGA_5$, we can see that engaging in similar image generation training further improves the performance of LLMGA in both generation and editing.

5 Conclusion

LLM possesses an extensive reservoir of knowledge and powerful comprehension and reasoning capabilities. In this paper, we introduce a MLLM-based generation assistant (LLMGA), aiming to exploit LLM's capabilities in an interactive manner to facilitate more efficient and convenient image generation and editing. Compared to relying on LLM to predict a fixed-size embedding to control SD, we employ detailed generation prompts. These prompts prove to be more favorable for enhancing LLM's contextual comprehension and generating more accurate and rich content. To this end, we develop a two-stage training scheme and curate a dataset, including four parts: prompt refinement, similar image generation, inpainting & outpainting, and instruction-based editing. For the first stage, we train MLLM to understand the properties of image generation and editing, enabling it to give fitting responses. For the second stage, we optimize the SD unet to adapt to the generation prompt. Moreover, we propose a DM-based reference restoration network (DiffRIR) to mitigate disparities in texture, contrast, and brightness for image editing. Consequently, LLMGA can offer design suggestions and enhance results based on user's requests during interactions.

Acknowledgements

This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS (2022)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. NeurIPS (2021)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: CVPR (2023)
5. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. arXiv preprint arXiv:2303.06555 (2023)
6. Batzolis, G., Stanczuk, J., Schönlieb, C.B., Etmann, C.: Conditional image generation with score-based diffusion models. arXiv preprint arXiv:2111.13606 (2021)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
9. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/
10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
11. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
12. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)
14. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
15. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)

16. Fan, W.C., Chen, Y.C., Chen, D., Cheng, Y., Yuan, L., Wang, Y.C.F.: Frido: Feature pyramid diffusion for complex scene image synthesis. In: AAAI (2023)
17. Feng, Z., Zhang, Z., Yu, X., Fang, Y., Li, L., Chen, X., Lu, Y., Liu, J., Yin, W., Feng, S., et al.: Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In: CVPR (2023)
18. Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023)
19. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: CVPR (2022)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
22. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)
23. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
24. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
25. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Direct inversion: Boosting diffusion-based editing with 3 lines of code. ICLR (2024)
26. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023)
27. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR (2022)
28. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. NeurIPS (2021)
29. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. arXiv preprint arXiv:2305.17216 (2023)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
34. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
35. Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
36. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
37. OpenAI: Gpt-4 technical report (2023)
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)

39. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. NeurIPS (2023)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
43. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
44. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH (2022)
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
47. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)
48. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023)
49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
50. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
51. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
52. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023)
53. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca` (2023)
54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
55. Valevski, D., Kalman, M., Matias, Y., Leviathan, Y.: Unitune: Text-driven image editing by fine tuning an image generation model on a single image. arXiv preprint arXiv:2210.09477 (2022)
56. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW (2021)
57. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)

58. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
59. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Timotfe, R., Van Gool, L.: Diffi2i: Efficient diffusion model for image-to-image translation. arXiv preprint arXiv:2308.13767 (2023)
60. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. ICCV (2023)
61. Xia, B., Zhang, Y., Wang, Y., Tian, Y., Yang, W., Timofte, R., Van Gool, L.: Knowledge distillation based degradation estimation for blind super-resolution. ICLR (2023)
62. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:2305.18295 (2023)
63. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
64. Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning. arXiv preprint arXiv:2309.02591 (2023)
65. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. NeurIPS (2024)
66. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
67. Zhong, S., Huang, Z., Wen, W., Qin, J., Lin, L.: Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In: ACM MM (2023)
68. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI (2017)
69. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)