An Empirical Study and Analysis of Text-to-Image Generation Using Large Language Model-Powered Textual Representation

Zhiyu Tan[1], Mengping Yang[1], Luozheng Qin[2], Hao Yang[1], Ye Qian[1], Qiang Zhou[1], Cheng Zhang[3], and Hao Li[4]

[1] INF Tech [2] Soochow University [3] Carnegie Mellon University [4] Fudan University

https://llm-conditioned-diffusion.github.io

Abstract. One critical prerequisite for faithful text-to-image generation is the accurate understanding of text inputs. Existing methods leverage the text encoder of the CLIP model to represent input prompts. However, the pre-trained CLIP model can merely encode English with a maximum token length of 77. Moreover, the model capacity of the text encoder from CLIP is relatively limited compared to Large Language Models (LLMs), which offer multilingual input, accommodate longer context, and achieve superior text representation. In this paper, we investigate LLMs as the text encoder to improve the language understanding in text-to-image generation. Unfortunately, training text-to-image generative model with LLMs from scratch demands significant computational resources and data. To this end, we introduce a three-stage training pipeline that effectively and efficiently integrates the existing text-to-image model with LLMs. Specifically, we propose a lightweight adapter that enables fast training of the text-to-image model using the textual representations from LLMs. Extensive experiments demonstrate that our model supports not only multilingual but also longer input context with superior image generation quality.

1 Introduction

The recent advancement of proficient text-to-image generative models, such as DALL·E series [4, 35, 36], Imagen [40], and Stable Diffusion [37], have sparked a revolution in creating various images conditioned on texts. Their technical breakthroughs not only push the boundaries of image synthesis but also significantly facilitate a spectrum of downstream applications, ranging from image editing and manipulation [19, 44, 59], personalized generation [3, 13, 38], and video generation [5]. In the context of text-to-image generation, a well-known fact is that the synthesis quality is greatly determined by the text features extracted from input prompts. The root reason is that the generated images are bounded by the text representation capabilities of the text encoders. Most existing text-to-image models [32, 35, 37] utilize CLIP [20, 33] to encode the input prompts. Although CLIP nudges language understanding for text-to-image diffusion models, it has three essential drawbacks: 1) CLIP is tailored for English only, non-English native users have to translate their prompts before using image synthesis models,

Fig. 1: Our proposed model could not only produce images with high visual quality given English input prompts (left), but also enables multilingual understanding capability for various language driven text-to-image generation (middle), as well as grasps much longer contextual information for generation (right).

which may introduce extra inference latency and lose some contextual information of the prompts; 2) the max token length of CLIP is limited to 77, resulting information loss in longer text conditions; and 3) the model capacity of CLIP is relatively small, leading to underperforming text representation capabilities, ultimately limiting the performance of text-to-image generative models.

In contrast, Large Language Models (LLMs) [1,2,6,31] have shown unprecedented progress and gained extensive attentions in both academic and industry communities recently. Attributed to ample training data, large-scale computation resources, LLMs herald a qualitative leap in text representation and language understanding abilities, which are remarkably better than those from CLIP models. Moreover, since LLMs are trained on multilingual corpora with a longer preserved context length, it also support multilingual inputs and longer context accommodation. These advantages indicate that incorporating LLMs into text-to-image diffusion models could effectively address the limitations of CLIP models and significantly bootstrap the synthesis quality.

However, employing LLMs as the text encoders of diffusion models is challenging. Unlike CLIP models that are trained to align features of image-text pairs, LLMs are developed solely from textual corpora, yielding text features that lack the awareness of visual information. Therefore, directly employing LLMs as the text encoders of diffusion models might be inappropriate due to the informa-

tion gap between textual and visual signals. Additionally, training text-to-image generative model with LLMs from scratch is resource-consuming, requiring massive computational and data resources [4, 7, 40]. For instance, DALL-E 3 [4] is trained on 1.2B images with a batchsize of 2048 for 500,000 iterations in total — a commitment of resources that many researchers and institutions cannot afford.

To address these issues, we propose a three-stage training pipeline that efficiently integrating LLMs into existing diffusion models. The main idea of our method is a lightweight but effective adapter module to align the text features of LLMs with that of the visual-aware CLIP. In this way, LLMs could capture the visual clues contained in the input prompts, thereby drive text-to-image diffusion models to produce appropriate images. Specifically, we decompose the training procedure into three distinct stages. First, we adapt the features of LLMs into diffusion training process by aligning them with those from CLIP models, only adapter is optimized in this stage. Then, we improve the synthesis quality through end-to-end text-image training. After that, the aesthetic appeal of the generated images is enhanced by further finetuning on a carefully-curated dataset. By doing so, the textual representation capabilities of LLMs can be fully activated and the model performance is well improved in terms of text alignment, synthesis quality and image aesthetics. Notably, our model is trained with a fraction of the resources required by most text-to-image diffusion models while achieving superior synthesis quality and supporting multilingual input.

In order to verify the effectiveness of our proposed model, we conduct extensive empirical investigation on both English and Chinese prompts datasets, it turns out our model achieves favourable zero-shot FID/CLIP scores under various settings. Besides, user studies demonstrate that our model could produce images that are preferred by human. In summary, our main contributions are:

- We propose an effective approach for incorporating LLMs into text-to-image diffusion models, improving the awareness of LLMs towards the CLIP visual and textual space, thus facilitating more expressive language understanding.
- We devise an efficient three-stage training pipeline that accomplish fast adaptation of LLM textual features with a small amount of resources, serving as an strong baseline of integrating LLMs into diffusion models and paving the way of this important topic.
- Extensive experiments demonstrate that our model not only supports multilingual input conditions (*i.e.,* Chinese, Japanese, Korean, *etc.*) but also surpasses existing text-to-image diffusion models in terms of synthesis quality, text alignment and image aesthetics.

2 Related Work

Text-to-image Diffusion Models. With the tremendous development of diffusion models [16, 29, 43], text-to-image generation [4, 32, 35, 37] have prompted appealing applications including image editing/manipulation [19, 44, 59], image translation/super-resolution [39, 50], personalized generation [13, 13, 38], *etc.* For

instance, given a text condition, DALLE-2 [35] employed the joint embedding space of CLIP [33] to learn a CLIP image embedding, based on which a decoder was trained to generate corresponding images. GLIDE [28] identified that leveraging classifier guidance [10, 17] for text-to-image generation brings better synthesis quality. Imagen [40] achieved better language understanding by leveraging a large transformer (*i.e.*, T5 [34]) language models to encode text conditions. Based on the text embedding power of CLIP [33], Latent Diffusion Model (LDM) [37], also know as Stable Diffusion (SD) [45], significantly improved the training efficiency of diffusion models by projecting samples into a pre-defined low-dimension latent space. Further, DALL·E-3 [4] and SD-XL [32] respectively improved text-to-image generation with better text captions, larger model capacity and novel conditioning schemes. Notably, SD-XL opted both OpenCLIP [20] and CLIP to obtain more powerful text understanding ability. Despite all these breakthroughs, most existing diffusion models are monolingual, *i.e.,* can only understanding one specific language, thus hindering their further applications. Moreover, the use of CLIP text encoder leads to very short contextual prompts understanding and underperforming expressive ability due to its maximum token length is only 77 and the model size of CLIP is relatively small.

Multilingual Text-to-image Generation. One straightforward way to produce multilingual images is to use machine translation tools to transform one language into another. For instance, when using ERNIE-ViLG [57] for text-to-image generation, one has to translate given prompts into Chinese before generation. This solution, however, is often time-consuming and leads to unsatisfactory results due to the expression, grammar, and cultural differences. To address this, several attempts have been made to enable multilingual diffusion models. The CogView series [11, 12] extracted a bilingual vocabulary of Chinese and English tokens to support both of them for text-to-image generation. Japanese SD [42] and Taiyi Diffusion [49] respectively extended the English-only SD by incorporating corresponding language knowledge to facilitate Chinese and Japanese text conditioned synthesis. Taiyi-Diffusion-XL [52] further improved the synthesis quality by integrating the capability of CLIP and SD-XL with bilingual continuous pre-training. In order to enable more various languages, AltDiffusion [55] first learned a multilingual text encoder and then re-train an initialized English-only diffusion model for concept alignment and quality improvement.

However, existing alternatives still have several limitations: First, training bilingual/multilingual text encoders from scratch usually requires massive training data and cost. Second, their language understanding abilities are still limited due to the misalignment between different languages and small text encoder model size. Third, there is often a trade-off between language understanding and image quality because more attention is paid to textual alignment and the more the text-to-image diffusion model is ignored. To address these issues, this paper proposes to leverage the text features of LLMs to provide promising language understanding ability and employ a resource-efficient three-stage training to ameliorate the synthesis quality, enabling a multilingual diffusion model without compromising the fidelity and diversity of generated images.

LLMs for Downstream Tasks. The recent advancement of LLMs [2, 6, 31] directly brings the blooming of visual content creation, multi-modal understanding, and various natural language processing tasks. For instance, the emergence of GPT family, namely GPT-3 [6], ChatGPT [30], GPT-4 [2], enabling convenient experiences for users to process tasks that were thought exclusive and burdensome before the era of LLMs. Motivated by their great success, researchers seek to tailor the power of LLMs for various downstream tasks, such as interpreting visual signals including images [51], videos [22], *etc.* For example, LLaVA [26,27] and MiniGPT-4 [60] utilized visual instruction learning to ameliorate the visual instruction following ability of LLMs. Video-LLaMA [58], ChatVideo [22] harnessed LLMs for video understanding tasks. These attempts either design novel projectors to unify the representation spaces between different modalities or align the embedding space of pre-trained feature extractors, one typical example is LLaVA that joint a vision encoder (*i.e.,* CLIP [33]) and an LLM (*i.e.,* Vicuna [8]) for multi-modal understanding. However, in the context of text-to-image generation, the potential of LLMs remain less-explored. Therefore, we present an effective practice of empowering text-to-image diffusion models with LLM-driven textual representation. By aligning the text embedding space of LLMs and the visual/textual embedding space of CLIP via a lightweight adapter, we propose a multilingual text-to-image diffusion model that enables longer contextual understanding and high-quality generation.

3 Method

In order to fully explore the potential of LLMs for text-to-image generation, we propose a novel three-stage training pipeline. First, a lightweight transformer-based adapter module is designed to align the text representations of LLMs with those of CLIP models, efficiently adapting the exceptional language understanding capabilities of LLMs for text-to-image generation. Then, an end-to-end training scheme is conducted to further optimize the adapter and the pretrained UNet, improving the synthesis quality. Finally, we perform a high-aesthetic finetuning on a small set of carefully curated high-quality images, improving the visual aesthetics of generated images. In the following, we first present necessary preliminaries in Sec. 3.1, followed by our overall framework in Sec. 3.2 and detailed description of our three-stage training pipeline in Sec. 3.3.

3.1 Preliminaries

Diffusion Models are trained to capture the distribution of training images through a sequential Markov chains of adding random noise into clean images and denoising pure noise to clean images. Formally, the forward process is accomplished by adding random Gaussian noise ($\epsilon \sim \mathcal{N}(0, \mathbf{I})$) into clean images from training sets ($x_0 \sim p_{data}$):

$$x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon, \tag{1}$$

where $t \in [0, 1]$ and $\gamma(t)$ is defined as a noise scheduler that monotonically descends from 1 to 0. On the contrary, the reverse process aims to denoise Gaussian noises back to clean images by iteratively predicting the added noises at each reverse step:

$$L(\theta) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathbf{I}),t} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t; t, c)\|^2 \right], \tag{2}$$

where ϵ_θ denotes the denoising model parameterized by a neural network, and c is the input conditions (*e.g.,* class condition or text condition).

Text Condition Injection in Diffusion Models. The most common way to inject text conditions into diffusion models is interacting text representations and with image features via cross-attention mechanism [32,37]. Concretely, image features are utilized as the query (Q), and textual features are deployed as both the key (K) and the value (V) within the cross-attention layer. The cross-attention layer computes a weighted sum of input sequences, where the weights are determined by the attention scores between Q and K: $\text{Attention}(Q, K, V) = \text{softmax}\left(QK^T/\sqrt{d_k}\right)V$. When using CLIP text encoder to inject textual conditions, the input text initially undergoes tokenization, yielding a sequence of N tokens. Note that if N exceeds a specified threshold T (77 for CLIP), the sequence will be truncated to T, and if N < T, the sequence will be padded to T. After that, the processed sequence is input into CLIP to extracts text feature, $h_{\text{clip}} \in \mathbb{R}^{T \times d}$, where d signifies the dimensionality. h_{clip} is then used to interact with visual features via cross-attention mechanism, determining the content of generated images with respect to the text conditions. It has been proved that the capacity of conditioned text feature plays a pivotal role in determining the quality of synthesized images [4,7,40]. Considering that LLMs [48,53] are trained on trillions of tokens with substantial model capacity for powerful language understanding, suggesting great potentials for text-to-image generation tasks.

3.2 Framework

Given the fact that the text features of LLMs are intrinsically ill-suited for text-to-image generation tasks, and directly training LLM-equipped diffusion models from scratch requires considerable resources, we derive an effective and efficient framework to adapt the textual representations of LLMs to diffusion models. Fig. 2 presents the overall framework. To be specific, we suffix an adapter module to the end of the LLM, achieving the alignment of text features between LLM and CLIP. The adapter is a 4-layer encoder-decoder transformer with a learnable query sequence q, which is defined as:

$$h_{\text{adapter}} := \theta_{\text{enc-dec}}(\theta_{\text{MLP}}(h_{\text{LLM}}), q), \tag{3}$$

where $h_{\text{LLM}} \in \mathbb{R}^{L \times d_{\text{LLM}}}$ and $h_{\text{adapter}} \in \mathbb{R}^{Q \times d_q}$, respectively denotes the output features of the LLM and the adapter. q is a hyper-parameter represents the number of learnable queries, which set as 77 in implementation.

The output text features of the adapter-suffixed LLM h_{adapter} serve as the text conditions of the diffusion training process, guiding the model to produce

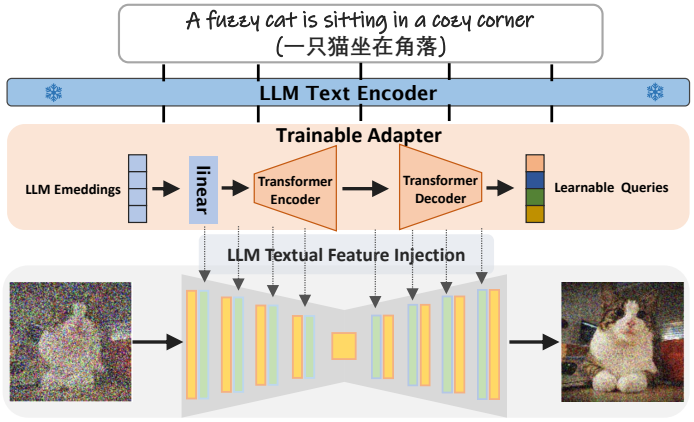


Fig. 2: Overall framework of our proposed method. The lightweight adapter efficiently connects LLMs and diffusion models, enhancing diffusion models with more powerful language understanding ability.

images that are consistent with input prompts. In practice, we first project initial images into a low-dimensional latent space as z_0 with a pre-trained VAE [32], and obtain noisy latent features z_t by adding noise ϵ to z_0. Similar to the Latent Diffusion Model (LDM) [37], we optimize our model by predicting the added noise conditioned on h_{adapter} at given timesteps t:

$$L_{\text{LDM}}(\theta) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, h_{\text{adapter}})\|^2\right], \tag{4}$$

where ϵ denotes the added noise and θ represents the model parameters.

3.3 Model Training Strategy

Our model is trained with three stages, namely multilingual textual alignment, end-to-end text-image training, and high-aesthetic finetuning. The overall pipeline is presented in Fig. 3. We describe details as follows.

Stage 1: Multilingual Textual Alignment. In order to introduce the text features of LLMs into the diffusion training process, we train the adapter to align the text feature of LLM with that of CLIP by merely optimizing the adapter through the proposed alignment loss. During training, we use two types of text corpora, English-only and Chinese-English paired text. For both of these datasets, the CLIP model is exclusively provided with English text, obtaining the features to be aligned, h_{clip}. Meanwhile, the adapter-suffixed LLM is respectively fed with English and Chinese text for English-only and Chinese-English corpora, yielding the adaption-needed LLM-derived features, h_{adapter}. Then, we aligning these two features by optimizing the adapter module through an alignment loss, which is defined as:

$$L(\theta_{\text{adapter}}) = 1 - \langle h_{\text{clip}}, h_{\text{adapter}} \rangle, \tag{5}$$

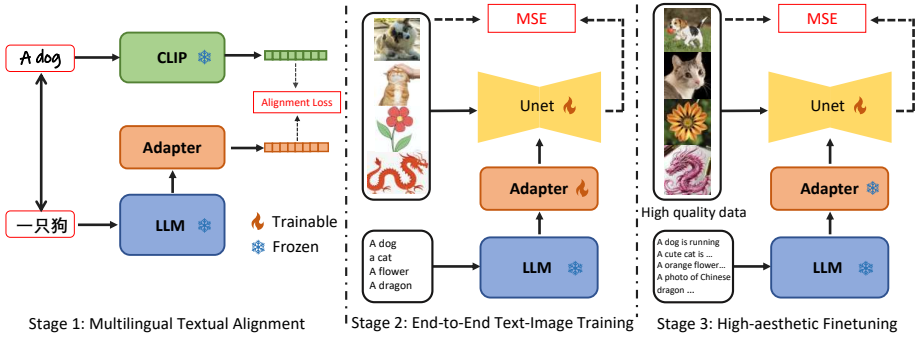


Fig. 3: Three-stage training pipeline. Multilingual textual alignment enables LLM to capture the visual and textual awareness of CLIP embedding space. End-to-end text-image training fully explores the potential of LLM-derived textual features and improves generation quality. High-aesthetic finetuning further ameliorates the visual aesthetic.

where $\langle\cdot\rangle$ is the cosine similarity. By minimizing the alignment loss, we can efficiently achieve multilingual textual alignment between LLM-derived features and CLIP image-text joint embedding space. Moreover, we observe that the magnitudes of the token-level features extracted from CLIP and LLM differ significantly, which impairs the alignment performance(detailed experiment is provided in Sec. 4.3). To address, we further incorporate a constraint on the token-level feature magnitudes of LLM and CLIP into the alignment loss:

$$L(\theta_{\text{adapter}}) = \lambda * (|h_{\text{clip}}| - |h_{\text{adapter}}|)^2 - \langle h_{\text{clip}}, h_{\text{adapter}} \rangle, \tag{6}$$

where the first term constrains the magnitude of these two features should be aligned, and the second term ensures the features from the adapter-suffixed LLM align with the feature space of the CLIP model, λ is the hyper-parameter that balance them.

Recall that the maximum token length of CLIP models is only 77, resulting information loss on text that is longer than 77. Therefore, directly aligning the adapter-suffixed LLM with the CLIP may bias the LLM-derived text features to short prompt text-to-image generation and hinders alignment performance. To mitigate this limitation, we adopted a segmented encoding approach. It first segments the input prompt into chunks of 77 tokens for separate CLIP encoding, the CLIP features of each chunk are then concatenated to obtain the complete feature. Such that, the information beyond 77 tokens are well-preserved without modifying the CLIP model, allowing multilingual textual alignment on longer text data.

Stage 2: End-to-end Text-image Training. Although stage 1 successfully aligns the textual feature space of the LLM to the CLIP embedding space, it does not fully explore the superior expressive capabilities of the LLM-derived text features, and the synthesis quality activated by LLMs could be further improved. Therefore, to enhance image synthesis quality and fully exploit LLM features,

Table 1: Quantitative comparisons between our proposed method and existing baselines. Our model achieves the best or second-best results in FID and CLIP-Score, while outperforms compared methods significantly in aesthetics score (Aes).

Model	COCO-en			COCO-cn			Flickr-30k-en			Flickr-8k-cn		
	FID	CLIP-s	Aes	FID	CLIP-s	Aes	FID	CLIP-s	Aes	FID	CLIP-s	Aes
Taiyi-Diffusion	**18.59**	0.2305	5.00	**72.58**	0.3389	4.98	80.66	0.2200	4.99	**78.00**	0.3412	4.98
Alt-Diffusion	33.20	0.2427	5.49	89.41	0.3382	5.47	83.73	**0.2522**	5.47	88.38	**0.3476**	5.48
Ours	22.38	**0.2463**	**6.80**	74.65	**0.3418**	**6.62**	**78.16**	0.2504	**7.03**	83.53	0.3326	**6.86**

this stage involves end-to-end text-image training on a curated internal dataset of 43M text-image pairs. During this stage, the parameters of LLM remain frozen, while the parameters of the adapter module and the UNet from the diffusion model are optimized by minimizing the diffusion loss, Eq. (2). Accordingly, the adapter is trained to further connect the textual representation space of LLMs and the text-to-image diffusion model, and the UNet is optimized to align the text conditions with the corresponding images, thus producing faithful and text-aligned images.

Stage 3: High-aesthetic Fine-tuning. Stage 2 ensures the alignment of image-text embeddings of LLMs and diffusion process, but the overall image quality is relatively random and unstable. In order to guide the model to exclusively produce images with highly visual aesthetics, we further finetune the diffusion model with a carefully curated high-quality images, which have shown appealing quality gains in the community [7,9]. Specifically, we filter 40K exceptionally high-quality images from a large amount of filtered images, where the filtering process includes text and optical character recognition, aesthetic score filter, image-text alignment, position, and object detection, as well as human visual judgment. Further, we employ a large multimodal model [27] to caption these images for highly descriptive text conditions with specific styles. In stage 3, the model is slightly optimized for nearly $5,000$ steps with a batchsize of 64.

4 Experiments

We conduct qualitative and quantitative experiments to validate our method. We first introduce experimental setups in Sec. 4.1. We then compare our methods with state-to-the-art models in Sec. 4.2. Finally, we show ablative studies and analyses in Sec. 4.3.

4.1 Experimental Setup

Datasets. Our three-stage training pipeline is respectively trained on pure text data (stage 1), text-image data (stage 2), and high-quality text-image data (stage 3). Regarding the text data, we use the text captions of LAION-5B [41] and collected real-world user prompts. For the paired text-image data for stage 2, we similarly employ the LAION-5B dataset and build an self-collected text-image

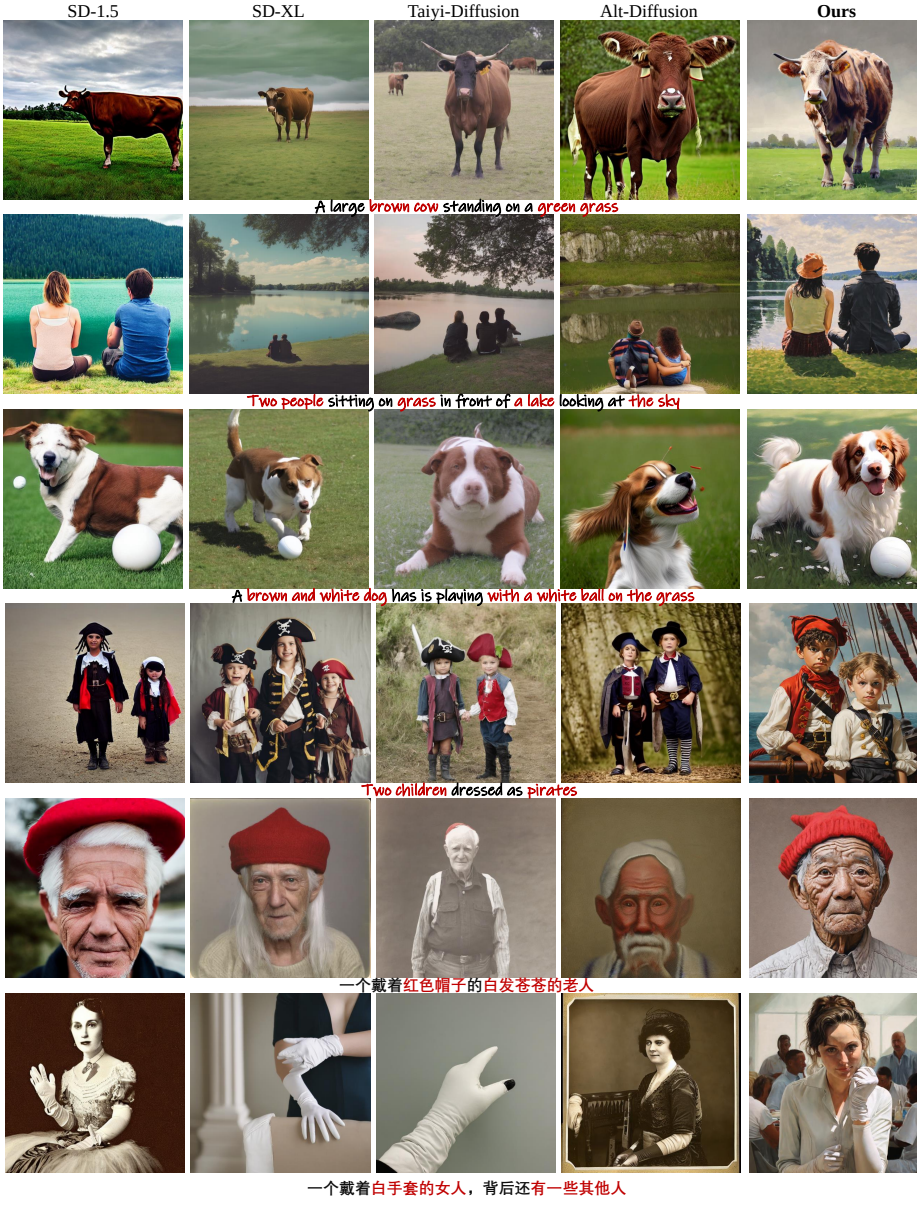


Fig. 4: Qualitative results of our model and competing methods. For models that do not support multilingual text conditions, we translate the given prompts into corresponding language to generate images. Our proposed method could produce images with accurate text-image alignment and higher visual quality.

dataset from internet, yielding a total of 43M samples. Notably, to fully exploit LLM-powered language understanding ability as done in [4, 7], we re-caption these images with a self-trained multi-modal LLMs, providing more descriptive text conditions for training. Finally, a carefully-crafted high-quality dataset is consist of both real-world images from internet and public-available image synthesis benchmarks [46], obtaining a total of 40K images with a average aesthetic score of 7.51. Notably, All of the aforementioned data is filtered, pre-processed and re-captioned with a carefully designed data cleaning process.

Evaluation Metrics. Following existing philosophy of evaluating text-to-image generative models, we include three commonly-used automatic metrics, CLIP-score [14], FID [15] and aesthetics score [41] to evaluate the generated images in terms of text alignment, synthesis quality and aesthetic appeal. Considering that our model features multilingual text conditions, we gather both English and Chinese prompts for a thorough investigation. Specifically, for English-driven text-to-image generation, we use COCO [25], Flickr-30K [56], and Flickr-8K [18], and use COCO-cn [24], Flickr-8K-cn [23] for Chinese driven text-to-image generation. Note that we only utilize the text captions for producing images in inference, and Chinese-CLIP [54] is used for calculating CLIP-Score of Chinese benchmarks. The original texts are used for computing CLIP-Score, and the original images are used as reference distribution for zero-shot FID evaluation.

To better investigate the visual quality of our proposed method, we further conduct an user study to identify the human preference of different models. Presented pairs of images generated by different models, 10 individuals are asked to select more visually appealing images. To ensure reliability of human evaluation, annotators are not aware of the image was produced by which model.

Comparison Baselines. Quantitatively, we compare our method with several existing multilingual text-to-image diffusion models, namely Alt-Diffusion [55], and Taiyi-Diffusion [49]. Moreover, for those baselines that do not support multilingual input conditions, *i.e.,* SD-v1.5 [45] and SD-XL [32], we compare their qualitative results based on machine-translated text conditions, *e.g.,* translating Flickr-8K-cn texts into English.

Implementation Details. We use the open-source Baichuan2-7B [53] as the integrated LLM, and build our 4-layer encoder-decoder adapter based on the SD-XL. All of the other details remain unchanged. The features aligned in stage 1 are text features concatenated from OpenCLIP and BigCLIP, similar to that of SD-XL. In stage 2, the UNet parameters are loaded from that of SD-XL and trained on image-text data with a batchsize of 3072, and classifier-free guidance is used by randomly dropping 10% input conditions. The whole training process is performed on 128 NVIDIA A800 GPUs for approximately 2 days.

4.2 Main Results

Quantitative Results. Tab. 1 presents the comparison results between our proposed method and existing alternatives. Owing to the aligned powerful language understanding of LLMs, our model achieves the best or the second-best

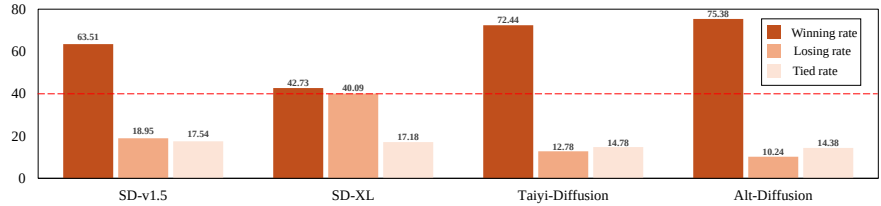


Fig. 5: Human evaluation results. Our model is consistently voted as the model that produce images with better visual quality.

FID/CLIP-Score performances under all tested benchmarks, while gains a significant improvement in visual aesthetics. It indicates that our model are able to generate images with competitive synthesis quality, textual alignment and aesthetic appeal.

Qualitative Results. Fig. 4 shows the visual outputs of our proposed method and comparison baselines, providing a clearer comparison on the synthesis quality. Clearly, the overall quality of images generated by our model is more visually appealing, and the prompts following and text-image alignment are significantly better than others. For instance, given "*A brown and white dog has is playing with a white ball on the grass*" as input, the image produced by our model is more realistic, the texture and visual details are more plausible. Moreover, in the last row, compared models miss generating other people in the back of the woman, and their output images look unnatural. This reflects that our model behaves better at following user prompts and synthesizing high-quality images.

Human Evaluation. In order to investigate the human preference of our model, we conduct user studies to compare with prior approaches in a pair-by-pair manner. Each time, we present users two images generated by our model and compared baselines, and then ask users to choose the better one. Fig. 5 shows the results. It turns out that our model is consistently identified as the better model when compared to different alternatives, particularly Taiyi-Diffusion and Alt-Diffusion. Together with the above quantitative and qualitative results, the advance of our proposed method is well proved.

4.3 Ablation Study and Further Analysis

Ablating Different Training Stages. Here we remove different stages of our proposed three-stages training pipeline to testify the efficacy of each stage. Specifically, we remove different training stages and remain other details unchanged, the quantitative results is shown in Tab. 2. These results demonstrate that: 1) stage 2 effectively improve the FID scores, suggesting its efficacy of ameliorating synthesis quality; 2) stage 3 greatly contributes to the image visual aesthetics scores; and 3) combining all stages could achieve the best trade-off in terms of visual quality, textual alignment, and aesthetics.

Ablating Magnitude Constraint. We added a magnitude to align the token of LLMs with CLIP text features, and here we provide the detailed analysis in

Table 2: Ablative results of different training stages.

Training Stages			COCO-en			COCO-cn			GPU Days
stage 1	stage 2	stage 3	FID	CLIP-s	Aes	FID	CLIP-s	Aes	
✓			22.73	0.2423	5.77	115.72	0.2157	6.20	18
	✓		15.50	0.2492	5.87	72.97	0.3311	5.49	53
✓	✓		16.92	0.2662	5.95	71.58	0.3560	5.50	53
✓		✓	44.11	0.2179	6.79	104.21	0.2731	6.78	18.2
✓	✓	✓	22.38	0.2463	6.80	74.65	0.3418	6.62	53.2

Fig. 6: Multilingual outputs. Our model is capable of understanding various languages and produce corresponding image content.

Tab. 3. It could be observed that after applying the magnitude constraint, the token differences between text features of CLIP and that of LLMs are effectively mitigated, and the CLIP-Score is improved. Such observations demonstrate the efficacy of explicitly aligning the encoded features of CLIP and LLMs.

Multilingual Text-to-image Generation. Fig. 6 provides the visual outputs of using various languages as the input conditions of our model. Surprisingly, our model could understand these texts well and generate images with corresponding captions. This amazing feature indicates that our model successfully integrates the powerful language understanding ability of LLMs into the text-to-image generation process, and fully exploit the potential of LLMs.

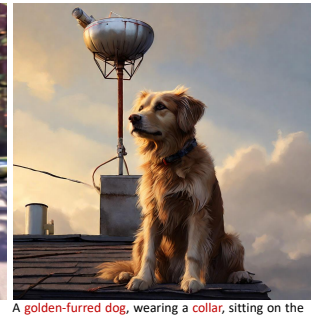
Long Prompts Generation. To testify whether our model could understand long texts, here we present the generated images under long prompts in Fig. 7. Remarkably, the proposed model could capture the meaning of prompts that are much longer than 77 tokens and synthesize images that well align with prompts,

Table 3: The magnitude differences between LLMs and CLIP text features

Words	A	majestic	lion	jumping	from	a	big	stone	CLIP-Score
CLIP	12.88	18.5	17.25	20.88	20.76	23.75	17.63	21.00	-
LLM w/ mc	14.44	18.23	18.14	21.87	21.86	20.76	21.03	20.84	0.2312
LLM w/o mc	17.94	17.6	21.77	18.77	18.14	15.74	15.15	14.64	0.2242

一位年轻女性，风格柔和且逼真。她长着一头深色的秀发，垂在肩上，皮肤白皙。她的五官精致，有着大而有神的眼睛，小巧的鼻子和丰满的嘴唇。她穿着一件印有花朵图案的上衣，与背景中的花朵相呼应。这件上衣有一个圆形领口，看起来是由轻薄、可能是丝质面料制成的。背景充满了粉色的花朵。光线柔和而温暖，给女性的脸和花朵投下了温柔的阴影。整个图像的氛围宁静而如画，唤起了宁静美丽感觉。

一位站在城市街道上的男士。他穿着一件浅色T恤，并留着胡须。他的目光直视镜头，看起来处于沉思或冥想的状态。男士位于前景，而城市景观在背景中模糊，表明了深度并聚焦于他。城市街道包含了交通灯、汽车和行人等元素，显示出是在一个城市环境中。光线和阴影暗示可能是傍晚或者初晚时分，太阳给场景投下了温暖的光芒。



A golden-furred dog, wearing a collar, sitting on the roof of a building next to a satellite dish and a surfboard. Its attention captivated by the distant view. The sky, painted with the soft hues of sunset, adds a sense of calm and tranquility to the scene. The image suggests a moment of quiet reflection or perhaps a watchful vigil.

Fig. 7: Long prompts driven generation results. Our model can capture long contextual text semantics and synthesize text-alignment results.

whereas prior methods usually fail under such setting. This further reflects the powerful language understanding capability and synthesis quality of our method.

5 Conclusion

In this paper, we present an empirical study on empowering text-to-image diffusion models with the exceptional language understanding capabilities of LLMs. We suffix a lightweight adapter to LLM to enable efficient feature alignment of LLM and CLIP, enabling text-to-image diffusion models to condition on LLM-derived text features. Specifically, we design a three-stage training pipeline to fully adapt the text features of LLM to the diffusion training process. Extensive results verify the effectiveness of the proposed training pipeline, yielding LLM-powered diffusion models that are able to create images with high synthesis quality, textual alignment and aesthetic appeal. Besides, our models also support text-to-image generation under multilingual and long-prompt scenario, serving as an effective and practical baseline for bootstrapping diffusion models with LLMs.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2, 5
3. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., H. Bermano, A.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 1
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. https://openai.com/dall-e-3 (2023) 1, 3, 4, 6, 11, 2
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 1
6. 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. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 5
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 3, 6, 9, 11
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org (2023) 5
9. 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) 9, 2
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 4
11. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in neural information processing systems **34**, 19822–19835 (2021) 4
12. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. Advances in Neural Information Processing Systems **35**, 16890–16902 (2022) 4
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. "International conference on learning representations" (2023) 1, 3
14. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 11, 2
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems (2017) 11, 2

16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 4
18. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of artificial intelligence research **47**, 853–899 (2013) 11, 5
19. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023) 1, 3
20. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773 1, 4
21. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 2
22. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 5
23. Li, X., Lan, W., Dong, J., Liu, H.: Adding chinese captions to images. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval. pp. 271–275 (2016) 11, 5
24. Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., Xu, J.: Coco-cn for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on multimedia **21**(9), 2347–2360 (2019) 11, 5
25. 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: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755 (2014) 11, 5
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 5
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 5, 9
28. 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) 4
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021) 3
30. OpenAI, T.: Chatgpt: Optimizing language models for dialogue. OpenAI (2022) 5
31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022) 2, 5
32. 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. arXiv preprint arXiv:2307.01952 (2023) 1, 3, 4, 6, 7, 11, 2
33. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 4, 5, 2

34. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The journal of machine learning research **21**(1), 5485–5551 (2020) 4
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 1, 3, 4
36. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. PMLR (2021) 1
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 3, 4, 6, 7
38. 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: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 22500–22510 (2023) 1, 3
39. 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 Conference Proceedings. pp. 1–10 (2022) 3
40. 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. Advances in neural information processing systems **35**, 36479–36494 (2022) 1, 3, 4, 6
41. 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. Advances in neural information processing systems **35**, 25278–25294 (2022) 9, 11, 2
42. Shing, M., Sawada, K.: Japanese stable diffusion. https://github.com/rinnakk/japanese-stable-diffusion (September 2022) 4
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International conference on learning representations (2021) 3
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International conference on learning representations (2021) 1, 3
45. Stability-AI: Stable diffusion public release (2023) 4, 11
46. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in neural information processing systems **36** (2024) 11
47. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 2
48. 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) 6
49. Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., Chen, C., Gan, R., Zhang, J.: Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. arXiv preprint arXiv:2209.02970 (2022) 4, 11

50. Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 7378–7387 (2023) 3
51. 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) 5
52. Wu, X., Zhang, D., Gan, R., Lu, J., Wu, Z., Sun, R., Zhang, J., Zhang, P., Song, Y.: Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support. arXiv preprint arXiv:2401.14688 (2024) 4, 6
53. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023) 6, 11
54. Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., Zhou, C.: Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335 (2022) 11
55. Ye, F., Liu, G., Wu, X., Wu, L.: Altdiffusion: A multilingual text-to-image diffusion model. arXiv preprint arXiv:2308.09991 (2023) 4, 11, 6
56. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the association for computational linguistics **2**, 67–78 (2014) 11
57. Zhang, H., Yin, W., Fang, Y., Li, L., Duan, B., Wu, Z., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. arXiv preprint arXiv:2112.15283 (2021) 4
58. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 5
59. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3836–3847 (2023) 1, 3
60. 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) 5

A Appendix

This appendix is organized as follows: First, we discuss the limitations and future works of the paper respectively in Appendix B and Appendix C. Next, we provide more analysis results to exhibit a more thorough perspective on the effectiveness of our approach in Appendix D. Specifically, we compare the synthesis performance of our proposed three-stage training pipeline under different alignment losses, including Mean Square Error(MSE), cosine similarity, and cosine similarity with Magnitude constraint. Additionally, trainable parameters are provided to demonstrate the efficiency of our proposed adapter. Then, we exhibit details of the human evaluation in Appendix E, including the user interface, user guidelines, and fine-grained results of user studies. Finally, we provide more qualitative results of multilingual and long-prompt text-to-image synthesis, as well as comparison results with existing baselines in Appendix F. Moreover, we provide the qualitative visual results synthesized by different stages of our training pipeline, together with the quantitative results (See Tab.2) in the main paper, the effectiveness of our three-stage training pipeline is well identified. Consistently, these qualitative results demonstrate that our proposed model effectively enables long-prompts following, multilingual, as well as better text-alignment understanding for text-to-image generation, providing users with satisfactory experience for producing visual content.

B Limitations

Despite its effectiveness of powerful language understanding capability and efficiency of training pipeline, our proposed model still have several limitations:

Human Evaluation. Conducting human evaluation can be time-consuming and resource-intensive, limiting the scale at which we can afford. In our case, 10 individuals with sufficient proficiency in both Chinese and English comprehension were involved to perform the human evaluation, each judged 400 image pairs produced by our proposed model and compared baselines. Although we had tried our best to lower the difficulty of the human evaluation by organizing the evaluation task as a pair-by-pair comparison way, the results of which may be inevitably biased. Additionally, human evaluation is inherently subjective and influenced by individual perspectives, biases, and preferences, particularly when it comes to the aspect of visual aesthetics. As a consequence, evaluation on a different set of prompts, or with different users and guidelines may lead to different result, thus making our evaluation somewhat limited.

Social Influence of Text-to-Image Models. Similar with other text-to-image models, our model can potentially generate images that contain offensive, inappropriate, or harmful content. If the input text includes biased or discriminatory language, the model might generate images that perpetuate stereotypes or promote harmful ideologies. Besides, our model may generate images that visually support false or misleading information provided in the prompt, potentially leading to the spread of misinformation or disinformation. We involved a robust and

unbiased dataset curation, created dedicated evaluation sets for bias detection and mitigation, and conducted adversarial testing through hours of redteaming, to alleviate these safety limitations. Moreover, we believe that with appropriate use, our proposed model could provide users with interesting experiences for content recreation, and inspires more appealing research works to connect LLMs with diffusion models.

Data-driven Finetuning. Following prior philosophy [4, 32], our model is trained on a large scale of image-text data to learn to fit the distribution of images conditioned on prompts. However, bounded by the quality and diversity of the training data, some parts of the distribution are badly fitted. Consequently, the model may struggle to generate certain objects that were not sufficiently learned during the data-driven finetuning. Furthermore, our approach might not fully bring the powerful language understanding capabilities of LLM towards text-to-image synthesis. Nevertheless, our approach strikes a considerable trade-off between efficiency and performance.

C Future Works

Besides scaling up our models in terms of training data and learnable parameters, we plan to expand our works from three perspectives. Firstly, we plan to try multiple adapter module and adopt the optimal one as the final module design of the adapter. Secondly, we plan to find a more reasonable automatic metric to re-assess our model. Thirdly, we plan to conduct quantitative evaluations on the multilingual text-to-image synthesis.

Although we have verified the effectiveness of our proposed adapter module, we simply utilized a 4-layer encoder-decoder transformer as the adapter module. Consequently, We leave the exploration for the optimal adapter module design for our further research.

Besides, throughout our experiments, we found it difficult to find a suitable automatic metrics to assess our models and planned to solve it in the foreseeable future. In our quantitative experiments, we utilized CLIP-score [14], FID [15], aesthetics score [41] to automatically evaluate the synthesis quality. Both the CLIP-score and FID do not correlate well with human assessment of the performance of generative models. This might be caused by CLIP [33] and Inception V3 [47], models used to calculate CLIP-score and FID. The CLIP and Inception V3 models are trained on low-resolution photos, which are significantly different from the image style preferred by our model. Meanwhile, the image-text consistency between the captions and ground truth images of popular benchmarks, are not so highly correlated that it further flawed the metric. The lack of correlation of existing automatic metrics is also noticed by the image synthesis communities [9, 21, 32].

Due to the lack of a widely adopted multilingual benchmark for text-to-image synthesis, we were unable to perform quantitative evaluations on the multilingual text-to-image synthesis capability of our model. In the future, we plan to deliver

a multilingual text-to-image synthesis benchmark and quantitatively verify the multilingual image synthesis ability possessed by our model.

D More Analysis Results

Analysis on Alignment Loss Variants. During the textual alignment stage, we conducted an ablation study on three variants of alignment loss functions: Mean Square Error (MSE) Loss, cosine similarity loss, and cosine similarity loss with magnitude constraint. Their formulations are illustrated below in Eq. (A1), Eq. (A2), and Eq. (A3), respectively.

$$L(\theta_{\text{adapter}})_{\text{MSE}} = (h_{\text{clip}} - h_{\text{adapter}})^2, \tag{A1}$$

$$L(\theta_{\text{adapter}})_{\text{cos}} = 1 - \langle h_{\text{clip}}, h_{\text{adapter}} \rangle, \tag{A2}$$

$$L(\theta_{\text{adapter}})^*_{\text{cos}} = 1 - \langle h_{\text{clip}}, h_{\text{adapter}} \rangle + (\text{mag}(h_{\text{clip}}) - \text{mag}(h_{\text{adapter}}))^2, \tag{A3}$$

where $\langle \cdot \rangle$ is the cosine similarity, $\text{mag}(\cdot)$ is the magnitude, $L(\theta_{\text{adapter}})_{cos^*}$ indicates the cosine similarity loss with magnitude constraint.

Specifically, we calculate CLIP-s on the validation set of the text alignment stage, the results of which are exhibited in Table A1. As can be concluded, models trained with cosine similarity loss perform better in terms of CLIP text-image similarity across three types of corpora compared to those trained with MSE loss. The English CLIP-S score improved from 0.1688 to 0.2318, and the Chinese CLIP-S score increased from 0.1923 to 0.2086. Furthermore, adding a magnitude constraint on top of the cosine similarity loss leads to even better alignment of the feature models.

Table A1: Ablation study results for different loss functions. "Baseline" indicates the original SD-XL model that integrated with CLIP text encoders, "Cosine Loss*" indicates our models optimized by cosine similarity loss with magnitude constraint.

Model	CLIP-s	CLIP-s(en)	CLIP-s(ch)
Baseline	0.2594	0.2868	0.2307
MSE Loss	0.1813	0.1688	0.1923
Cosine Loss	0.2242	0.2318	0.2086
Cosine Loss*	0.2312	0.2482	0.2168

Analysis on Module Parameters. To quantitatively investigate the extra calculating budget brought by the adapter module, we summarize the parameter numbers of each module utilized in our model, and illustrated them in Table A2. Our model contains nearly 10.6B parameters, a substantial parameter number endows significant fitting ability on the distribution of high quality and aesthetic

Table A2: Module Parameter Comparison. "#Parameters" and "Percentage" respectively denote the number of module parameters and module parameter contribution percentage in model, respectively.

Module	Our Adapter	VAE	UNet	LLM	Total
#Parameters	487M	83M	2.5B	7.5B	10.6B
Percentage (%)	4.6	0.8	24.1	70.5	100.0

images conditioned on prompts. The main contributor to the model parameters are the LLM text encoder (7.B) and the UNet (2.5B), together accounting for 94.6% of the total number of parameters in the model. Meanwhile, the extra introduced adapter is 487M, contributing merely 4.6% of parameters in our model. Therefore, the adapter module integrated in our model is quite lightweight, and the aforementioned quantitative experiments has confirmed its effectiveness.

E Human Evaluation

Fig. A1: Demonstration of our user interface. Each time, our specially designed user interface will provide a pair of images and the prompt used to generate them to the users. We incorporated six distinct icons to signify various functionalities of the user interface.

The human evaluation was conducted in a pair-by-pair manner, with each user required to compare 400 identical image pairs.

Data Preparation To prepare the data utilized for human evaluation, we first sample 400 prompts for image synthesis. Specifically, each baseline model is al-

located an equal distribution of 100 prompts, consisting of 50 Chinese prompts sampled from COCO-cn [24] and Flickr-8K-cn [23], and 50 English prompts sampled from COCO [25], Flickr-30K [18], and Flickr-8K [18]. Then, the baseline models generate images guided by their respective prompts, while our models generate an image conditioned on the same prompt. Subsequently, we create pairs between the images conditioned on the same prompt for comparison. Overall, we prepared 400 pairs of images for human evaluation, with each pair composed of two images generated with the guidance of the same prompt, one by our model and the other one by baseline models.

User Interface and Evaluation Procedure. To accomplish the evaluation effectively and efficiently, we designed an easy-to-understand user interface for users. Figure A1 demonstrates the functionalities of our user platform. Each time, users are provided with a prompt and a pair of generated images. Typically, users are required to click one of the two "like" buttons below each image, indicating their preference on a specific image. If users are unable to distinguish a better image, they can click the red X button centrally below the images. If users need to revise their previous compared pair, they can click the blue left and right arrow to jump to the previous and the next pairs. In cases that users are confused about the evaluation, they can utilize the raise hand button to ask the help of the organizers. Thanks to our adequate user training, reasonable evaluation procedure and friendly user interface, the raise hand button is set but not used in our evaluation. Notably, the image pairs are disordered, thus users are blind to the models that generates the comparing images at the time.

User Training and Guidelines Before performing human evaluation, we made a series of efforts to guarantee it to be fair and unbiased. First of all, we conduct a comprehensive user selection to select users that are willing to tolerate any inconvenient visual content that might be generated by the evaluated models. Generally, 10 out of 33 users were selected to perform the evaluation, consisting of 3 graduate students major in text-to-image generation, three artist with sensitive to aesthetics, and four amateurs. Then, to make users aware of the purpose and the standard for the human evaluation, we conduct a training over our carefully prepared user guidelines. As exhibited in Table A4, the user guidelines contain including instructions of user interface and detailed guidelines of human evaluation. During the training, each user is provided with a copy of the user guidelines in Chinese, and is asked to understand these user guidelines. We also explained to users the general purpose and standard for the human evaluation, and answer their questions about the user guidelines. Notably, in cases that users might be confused about the compared pairs in progress, we add a raise hand button in the user interface so that users are able to ask for our help at any time. Thanks to our adequate user training, reasonable evaluation procedure and friendly user interface, the raise hand button is set but not used.

Fine-grained Results of Human Evaluation. We illustrate fine-grained results of human evaluation in Table A3. As can be seen, our human users reached agreement that our model obviously beats SD-v1.5, Taiyi-Diffusion and Alt-

Table A3: Fine-grained Results of User Studies. "Taiyi-Diff" and "Alt-Diff" are abbreviation for "Taiyi-Diffusion" [52] and "Alt-Diffusion" [55], respectively. Our model gains much more votes by all participants when compared with various baselines, especially compared with multilingual models, *i.e.,* Taiyi-Diffusion and Alt-Diffusion.

User ID	*v.s* **SD-v1.5**			*v.s* **SD-XL**			*v.s* **Taiyi-Diff**			*v.s* **Alt-Diff**		
	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie
No.1	36	34	30	25	37	38	52	20	28	59	16	25
No.2	75	15	11	56	42	2	84	12	4	84	9	7
No.3	67	25	8	40	51	9	71	22	7	88	6	6
No.4	41	21	38	34	44	22	49	15	36	54	16	30
No.5	66	9	25	54	24	22	76	6	18	79	3	18
No.6	77	17	6	42	46	12	80	12	8	77	15	8
No.7	72	13	16	54	26	20	86	3	11	87	2	11
No.8	66	29	5	32	52	16	77	12	11	86	12	2
No.9	69	17	14	53	33	14	77	13	10	78	9	13
No.10	64	20	16	43	39	18	78	11	11	83	10	7

Diffusion. As for SD-XL, the majority of the ten users agreed that the image generated by our model is slightly better than that by SD-XL.

F More Quantitative Results

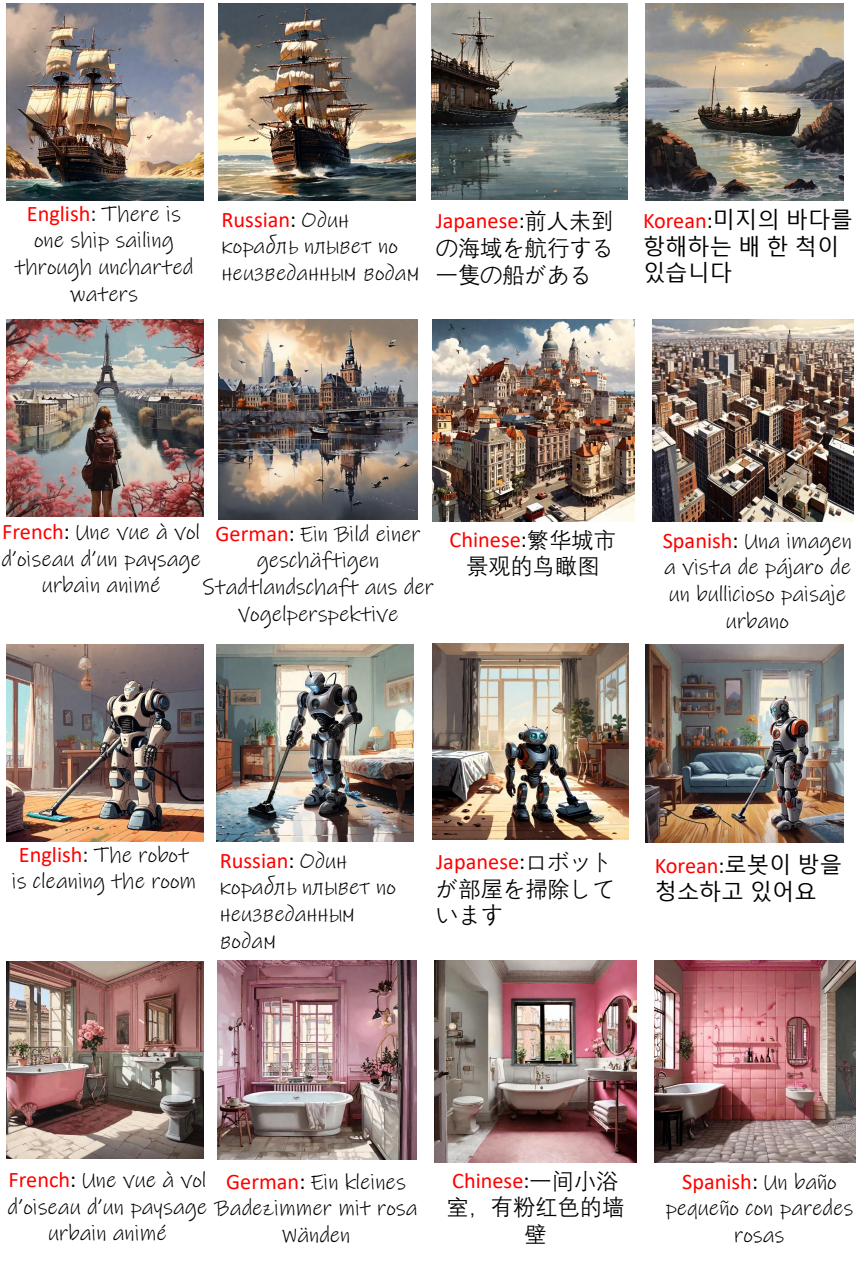


Fig. A2: Multilingual outputs. More examples showcasing that our model is capable of understanding various languages and produce corresponding image content.

Table A4: User Guidelines of the Human Evaluation. Considering that our users are Chinese speakers while our readers may not be, each user is provided with a copy of Chinese version of the user guidelines. Meanwhile, we demonstrate its translated English version in the following.

User Guidelines
Part I Guidelines of the User Interface
1.You are supposed to compare 400 pairs of images on the provided user interface throughout the whole evaluation process. 2.Each time, the user interface will show you a sequential number, a pair of images generated by different models, a prompt used to guide the generation of the images, and six different buttons to signify various functionalities of the user interface. 2.The sequential number for the comparing pairs printed at the top left corner of the user interface, is used to remind you of your working progress. 3.To accomplish the comparison of presenting image pairs, you are supposed to click on the "like" icon below a specific image. It indicates your preference for that specific image. 4.In cases where you likes or dislikes both images, you can click the red "X" button located centrally below the images. 5.Once you perform guideline 3 or guideline 4, the user interface will automatically switch to the next pair. 6.In the bottom right corner of the interface, there are a left and a right arrow icons, which enable you to navigate to the previous and next image pairs and revise your comparison results. 7.If you are unsure of anything about the human evaluation, feel free to click the raise hand button in the bottom right corner of the user interface, we will be glad to help you.
Part II Guidelines of the Human Evaluation
1.In general, you are supposed to compare the image pairs with the criterion of your personal preference. 2.If you are unsure about how to form your personal preference on the compared images. We suggest three criterion, image quality, text-image consistency and aesthetics. 3.Image Quality: whether the generated image is clear, details of the generated image objects are consistent with that in real world. 4.Text-image Consistency: whether the generated image is consistent with the description of the given prompt. 5.Aesthetics refers to the overall aesthetic quality of a generated image. It combines various visual elements such as color, shape, texture, and composition that creates a fascinating image as a whole. 6.You are required to first compare 30 pairs of images to form a stable and reasonable assessment standard. Then, compare the whole 400 pairs of images from start to finish. 7.If you feel confused at anything about the human evaluation, feel free to click the raise hand button in the bottom right corner of the user interface, we will be glad to help you. 8. If you finished the comparison of all 400 pairs, you can submit your evaluation results by clicking the right arrow button. 9. Once you have submitted your evaluation results, we are very thankful to inform you that you have finished your job. Thank you once again for your contribution to our project.

Near the village, the crystal-clear water gently laps against the shore. Beautiful and colorful houses line the coastline, creating a picturesque scene that seems straight out of a postcard. A fisherman is about to navigate his boat through the tranquil waters, searching of a bountiful catch. People are strolling along the waterfront and enjoying the serene atmosphere, as birds gracefully soar overhead, adding to the idyllic ambiance.

On the tranquil balcony, there is a glass jar takes its place. Inside, five Mars astronauts stand tall, clad in their space gear, their gazes fixed upon a sailing ship that emerges on the distant horizon. As the sun begins its descent, casting a warm golden hue across the sky, a surreal ambiance settles over the scene. Besides, as fish gracefully soar through the sky, the surreal ambiance is deepened. The fading daylight casts long shadows, evoking tranquility and excitement.

In the depths of the forest, a fearless chimpanzee explorer roams, donning an unzipped jacket that reveals its chiseled abdominal muscles, exuding an aura of immense strength. With a clenched fist, it locks its gaze upon the camera, unaware of the imminent threat posed by two menacing birds plotting their attack.

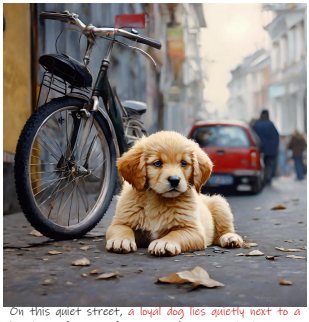
在这个极富未来科技感、干净整洁的餐厅中庭，光线柔和地洒落在整齐排放的散发出沉稳与传统气息的红木桌椅上。这些红木桌椅，与餐厅天花板上的霓虹灯形成鲜明对比，后者象征着科技与喧闹。而修剪整齐的灌木则静静地点缀着角落，为人们带来一份宁静和放松，使人们感受到与大自然的联系。人们能够在这里沉浸于科技与自然、传统与前卫的双重魅力之中，感受到一份全新的美好体验。

在这热闹的小镇街道上，行人和各种载具穿梭而过，营造出一片熙熙攘攘的景象。其中一条街区旁停放着一辆不起眼的蓝色汽车，静静地等待着主人的归来。在小镇的远方，依稀可见高楼大厦直冲云霄，矗立在小镇的天际线上，与相比较为传统的小镇建筑形成对比。它们的外观精美，为小镇注入了一丝科幻和现代化的色彩。与此同时，一艘UFO形状的飞行器掠过小镇，划破了宁静的天空。

在海底的酒吧中，玻璃般的墙壁波光粼粼，创造出一副迷人的景象。玻璃墙壁外，鱼儿优雅地在水中游动。而在酒吧内，顾客们聚集在吧台旁，一边品酒，一边欣赏着周围的水下美景。他们惊叹地观赏着鱼儿在珊瑚礁间穿梭游动，随着海水的起伏翩翩起舞。酒吧如玻璃般的墙壁提供了一个无障碍的视野，让顾客们感觉自己仿佛融入了水下世界。吧台上摆放着众多酒瓶，诱人的酒香让顾客们沉醉其中。

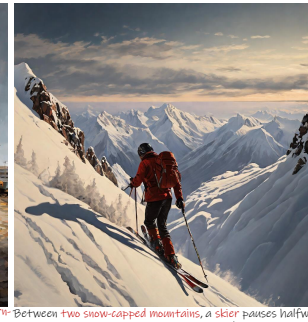
A group of travelers is visiting a massive golden Buddha statue situated on a high platform. Reaching the platform, the group was greeted by the resplendent sight of the towering golden statue. Its majestic stature commanded attention, standing tall and dignified. The intricate details of the statue's design were masterfully crafted, with every curve and contour reflecting a profound sense of spirituality and inner peace.
As the sunlight bathed the golden Buddha, its radiance seemed to merge with the surrounding clouds, creating an ethereal ambiance. The interplay of light and shadow cast a mesmerizing glow that transformed the statue into a beacon of transcendence. The travelers stood in silent awe, their hearts filled with a deep sense of reverence and wonder.

Fig. A3: Long prompts driven generation results. More examples showcasing that our model could capture long contextual text semantics and synthesize text-alignment results. Zoom in for details.



On this quiet street, a loyal dog lies quietly next to a bicycle, as if waiting for its owner's return. The houses on either side of the street stand tall, and a gentle breeze carries the refreshing scent of flowers, creating a peaceful and harmonious atmosphere. Behind the dog, a red car faces away from them, gleaming in the sunlight, forming a striking contrast with the surrounding scenery.

Under the gloomy sky, an elderly man sits quietly on a worn-out stool by the harbor, holding a pale yellow umbrella in his hand. His gaze is deep and tranquil, as if staring at the distant city shrouded in dark clouds. In the distance, a massive cargo ship moves slowly. Ripples appear on the sea's surface, seemingly narrating the vicissitudes and vastness of time. This scene is like an oil painting, blending silence, solitude, and warmth together.

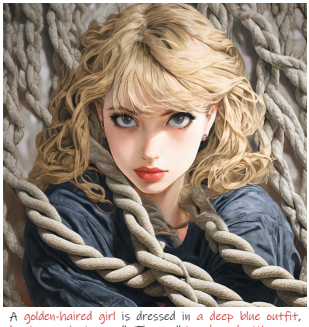


Between two snow-capped mountains, a skier pauses halfway down the slope, clad in vibrant skiing gear. He stands amidst the pristine white snow, as if a part of nature, blending into the serenity and grandeur of this snowy realm. The towering snow-capped mountains on both sides glow with a golden hue in the sunlight. The distant mountain range undulates, appearing ethereal and elusive in the evening's glow, showcasing a dreamlike beauty and mystique.

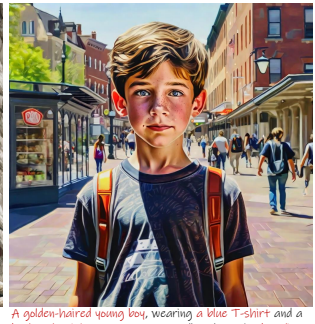
在城市的中心，一辆古老的有轨电车悠然驶过，仿佛是时间的见证者。行人匆匆忙忙，如潮水般涌动，来自不同方向，去向不同目的地。高楼拔地而起，宛如巨人般向天空拥抱，反射出耀眼光芒，构成一幅现代城市的画卷。这个画面充满着都市的生机与活力。有轨电车的轨道延伸至远方，连接起城市的过去与现在，见证着岁月的更迭。在这个画面中，历史与现代相互交织，高楼与有轨电车共同构成了城市的风景。

在一艘没有帆的小船上，一位孤独的人静静地坐着，凝视着远处的大帆船和天空中太阳透过云层洒下的金光。这艘无帆的船静静漂浮在宁静的海面上，仿佛是时间的凝固，承载着无尽的孤寂和思绪。远处，一艘巨大的大帆船如同一只自由飞翔的鸟儿，航行在辽阔的海洋上。太阳从云层中探出，洒下一道金色的光芒，照耀着大海和船帆，映照出迷人的光影变幻。

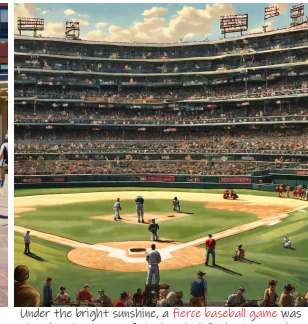
在一间充满电子元件的房间里，一个年轻女性坐在一堆电子元件前，专注地处理着手头的工作。她眉头微蹙，手指灵巧地操作着微小的电子零件，仿佛在进行一场精密的舞蹈。在她背后的墙上，密密麻麻挂满了连绵不断的电线和各种电子设备。电线交错缠绕，形成一幅错综复杂的画面，就像是大脑中神经元的连接方式，充满了生机与活力。



A golden-haired girl is dressed in a deep blue outfit, leaning against a wall. The wall is adorned with many rough hemp ropes hanging down from the top, gently swaying in front of the girl, as if following a mysterious melody. Her clear and bright eyes reveal a wisdom and maturity that surpass her age.



A golden-haired young boy, wearing a blue T-shirt and a backpack with orange straps, walks along the bustling street. He faces the camera, with the sunlight shining on his back, exuding a vibrant sense of youth. His eyes sparkle with a youthful enthusiasm, eagerly observing the bustling surroundings. Despite the busy street, he remains unfazed, radiating a sense of calm and self-assurance.



Under the bright sunshine, a fierce baseball game was played in the center of the baseball field. The stands were packed, and the audience was enthusiastic. Cheers and shouts came and went, floating in the air like musical notes. The surrounding environment also adds a different kind of charm, with a few white clouds floating in the sky, shining sunlight onto the court, illuminating every detail.

Fig. A4: Long prompts driven generation results. More examples showcasing that our model could capture long contextual text semantics and synthesize text-alignment results. Zoom in for details.

Fig. A5: Comparison results between stages. "Stage i" indicates the images generated by models trained after Stage i, where i∈[1, 2, 3]. Clearly, the synthesis quality, including the overall image structure, fine-details, visual aesthetics improves with the training stages progresses.

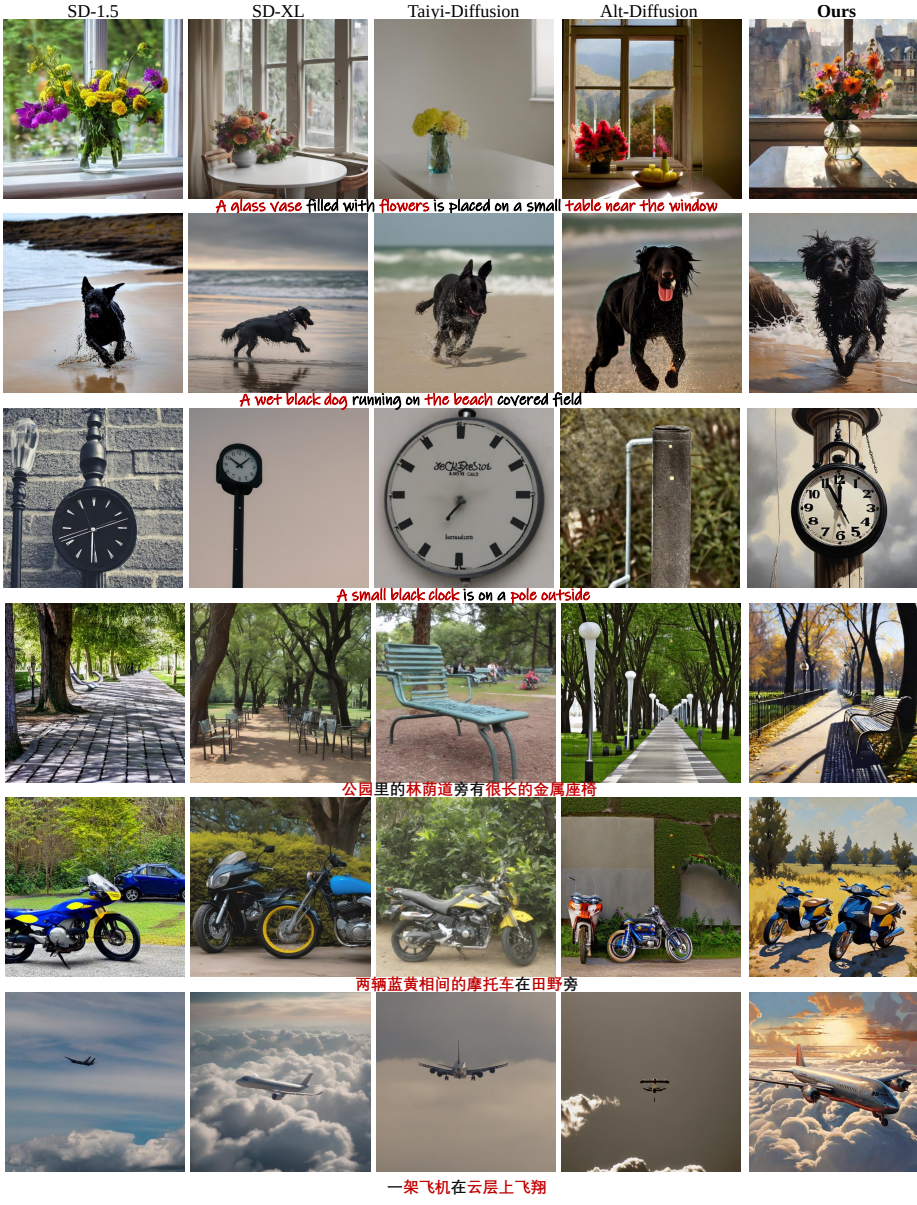


Fig. A6: More visual comparison results between our proposed method and popular text-to-image diffusion baselines. For models that do not support multilingual text conditions, we translate given prompts into corresponding language to generate images for comparison. Our proposed method could produce images that are plausible, with better text-image alignment and higher visual quality.

Fig. A7: More visual comparison results between our proposed method and popular text-to-image diffusion baselines. For models that do not support multilingual text conditions, we translate given prompts into corresponding language to generate images for comparison. Our proposed method could produce images that are plausible, with better text-image alignment and higher visual quality.