Kolors: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis

Kolors Team
Kuaishou Technology
kwai-kolors@kuaishou.com

Code: https://github.com/Kwai-Kolors/Kolors

Weights: https://huggingface.co/Kwai-Kolors/Kolors

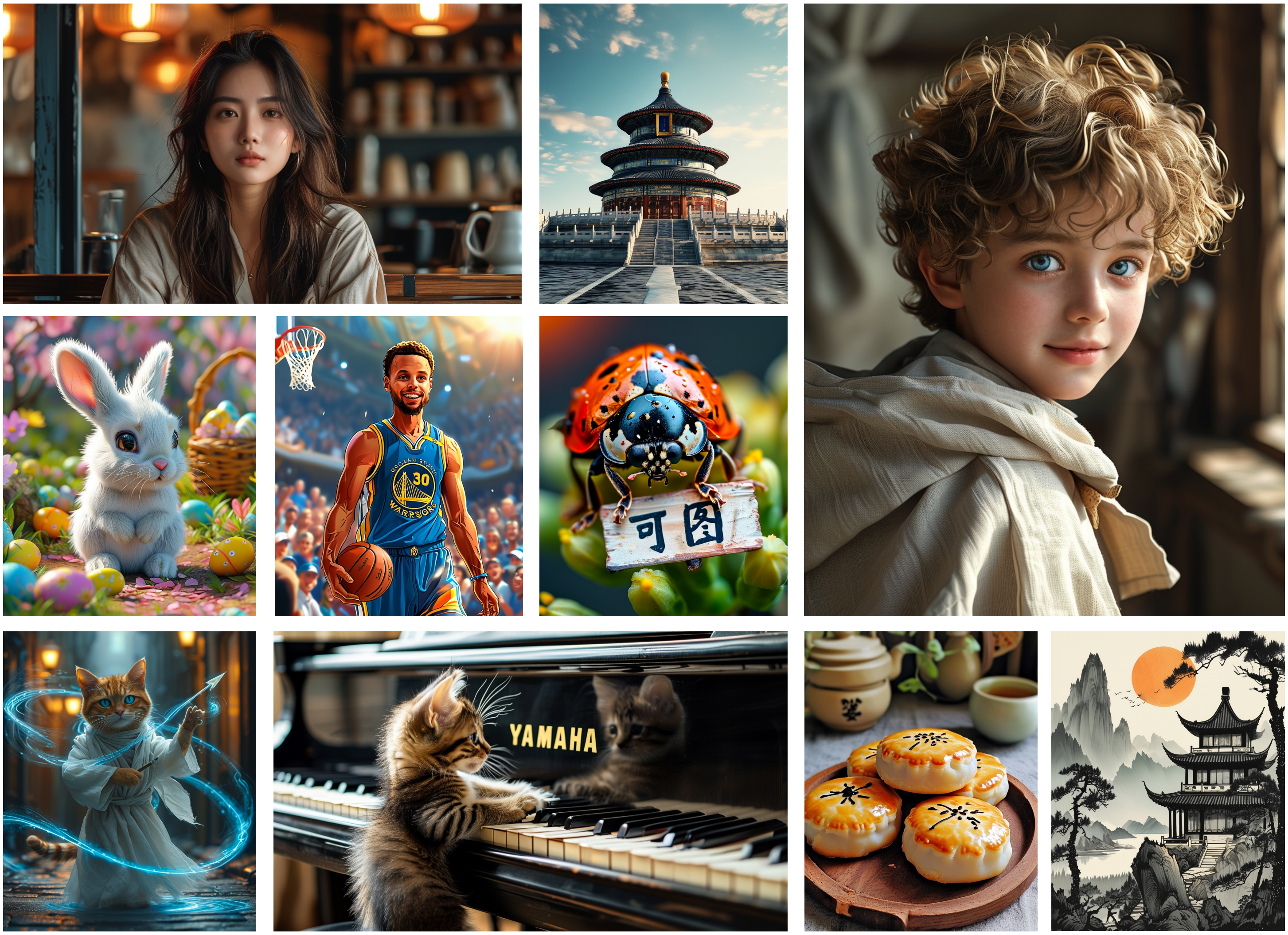


Figure 1: High-quality samples generated by Kolors, showcasing its outstanding capabilities in English and Chinese text rendering, precise prompt adherence, meticulous detail rendering, and superior image quality across a wide range of styles and resolutions.

Abstract

We present ***Kolors***, a latent diffusion model for text-to-image synthesis, characterized by its profound understanding of both English and Chinese, as well as an impressive degree of photorealism. There are three key insights contributing to the development of Kolors. Firstly, unlike large language model T5 used in Imagen and Stable Diffusion 3, Kolors is built upon the General Language Model (GLM), which enhances its comprehension capabilities in both English and Chinese. Moreover, we employ a multimodal large language model to recaption the extensive training dataset for fine-grained text understanding. These strategies significantly improve Kolors' ability to comprehend intricate semantics, particularly those involving multiple entities, and enable its advanced text rendering capabilities. Secondly, we divide the training of Kolors into two phases: the concept learning phase with broad knowledge and the quality improvement phase with specifically curated high-aesthetic data. Furthermore, we investigate the critical role of the noise schedule and introduce a novel schedule to optimize high-resolution image generation. These strategies collectively enhance the visual appeal of the generated high-resolution images. Lastly, we propose a category-balanced benchmark *KolorsPrompts*, which serves as a guide for the training and evaluation of Kolors. Consequently, even when employing the commonly used U-Net backbone, Kolors has demonstrated remarkable performance in human evaluations, surpassing the existing open-source models and achieving Midjourney-v6 level performance, especially in terms of visual appeal. We will release the code and weights of Kolors at `https://github.com/Kwai-Kolors/Kolors`, and hope that it will benefit future research and applications in the visual generation community.

1 Introduction

Diffusion-based text-to-image (T2I) generative models have emerged as focal points in the artificial intelligence and computer vision fields. Previous methods, such as SDXL [27] by Stability AI, Imagen [34] from Google, and Emu [6] from Meta, are built on the U-Net [33] architecture, achieving notable progress in text-to-image generation tasks. Recently, some Transformer-based models, such as PixArt-α [5] and Stable Diffusion 3 [9], have showcased the capability to generate images with unprecedented quality. However, these models are currently unable to directly interpret Chinese prompts, thereby limiting their applicability in generating images from Chinese text. To improve the comprehension of Chinese prompts, several models have been introduced, including AltDiffusion [45], PAI-Diffusion [39], Taiyi-XL [42], and Hunyuan-DiT [19]. These approaches still rely on CLIP for Chinese text encoding. Nevertheless, there is still considerable room for enhancement in terms of Chinese text adherence and image aesthetic quality in these models.

In this report, we present Kolors, a diffusion model incorporating the classic U-Net architecture [27] with the General Language Model (GLM) [8] in the latent space [32] for text-to-image synthesis. By integrating GLM with the fine-grained captions produced by a multimodal large language model, Kolors exhibits an advanced comprehension of both English and Chinese, as well as its superior text rendering capabilities. Owing to a meticulously designed two-phase training strategy, Kolors demonstrates its remarkable photorealistic capabilities. Human evaluations on our ***KolorsPrompts*** benchmark have confirmed that Kolors achieves advanced performance, particularly excelling in visual appeal. We will release the code and model weights of Kolors, aiming to establish it as the mainstream diffusion model. The primary contributions of this work are summarized as follows:

- We select GLM as the appropriate large language model for text representation in both English and Chinese within Kolors. Furthermore, we enhance the training images with detailed descriptions generated by a multimodal large language model. Consequently, Kolors exhibits exceptional proficiency in comprehending complex semantics, particularly in scenarios involving multiple entities, and demonstrates superior text rendering capabilities.
- Kolors is trained with a two-phase approach that includes the concept learning phase, using broad knowledge, and the quality improvement phase, utilizing carefully curated high-aesthetic data. Furthermore, we introduce a novel schedule to optimize high-resolution

image generation. These strategies effectively improve the visual appeal of the generated high-resolution images.

- In comprehensive human evaluations on our category-balanced benchmark, ***KolorsPrompts***, Kolors outperforms the majority of both open-source and closed-source models, including Stable Diffusion 3 [9], DALL-E 3 [3], and Playground-v2.5 [18], and demonstrates performance comparable to Midjourney-v6.

2 Methods

In this work, we concentrate on several critical aspects to enhance the performance of text-to-image diffusion models. Although optimizing a more advanced backbone could potentially improve model performance, it falls outside the scope of this study. For the backbone, we strictly adhere to the U-Net architecture as utilized in SDXL [27]. Importantly, our methods are effective across various diffusion models, not limited to the latent U-Net [33].

2.1 Enhancing Text Faithfulness

2.1.1 Large Language Model as Text Encoder

The text encoder is a crucial component of text-to-image generative models, directly influencing and controlling the content generated by the model. We compare the text encoder usage of typical image generation models with our Kolors in Table 1. Generally, the CLIP [28] and T5 [29] series dominate as text encoders. Classic methods, such as SD1.5 [32] and DALL-E 2 [30], employ the text branch of the CLIP model for text representation. However, since CLIP is trained with a contrastive loss to align entire images with text descriptions, it struggles to understand detailed image descriptions involving multiple subjects, positions, or colors.

Some methods extract text embeddings from an encoder-decoder Transformer T5, which carries more fine-grained local information, such as Imagen [34] and PixArt-α [5]. Additionally, other methods utilize multiple text encoders for enhanced text understanding. For instance, eDiff-I [2] proposes an integrated text encoder combining CLIP and T5 for both global and local text representation. SDXL [27] employs two CLIP encoders and achieves promising results in the open-source community. Stable Diffusion 3 [9] further incorporates a T5-XXL text encoder into its model architecture, which is essential for handling complex prompts. Recently, LuminaT2X proposes a unified framework to transform text into any modality by leveraging the pre-trained LLM model LLama2 [38].

Notably, most text-to-image generative models encounter difficulties with Chinese prompts due to the limitations of the English text encoders in CLIP. HunyuanDiT [19] addresses this issue by employing a bilingual CLIP and a multilingual T5 [43] encoder for Chinese text-to-image generation. However, the training corpus for Chinese texts constitutes less than 2% of the multilingual T5's dataset, and the text embeddings produced by the bilingual CLIP are still inadequate for handling complex text prompts.

To address these limitations, we choose the General Language Model (GLM) [8] as the text encoder in Kolors. GLM is a bilingual (English and Chinese) pre-trained language model based on an autoregressive blank infilling objective, which significantly outperforms both BERT and T5 in natural language understanding and generation tasks. We posit that the pre-trained ChatGLM3-6B-Base model is more suitable for text representation, whereas the chat model ChatGLM3-6B, which has undergone human preference alignment training, excels in text rendering. Consequently, in Kolors, we utilize the open-source ChatGLM3-6B-Base as text encoder, which has been pre-trained with over 1.4 trillion bilingual tokens, resulting in a robust capability for Chinese language understanding.

Compared with 77 tokens of CLIP, we directly set the text length of ChatGLM3 up to 256 for detailed complex text understanding. Note that it is easy to support long text, due to the long text processing capability of ChatGLM3. Following SDXL, the penultimate output of ChatGLM3 is used for text representation.

2.1.2 Improved Detailed Caption with Multimodal Large Language Model

The training text-image pairs are often sourced from the Internet, and the accompanying image captions are inevitably noisy and inaccurate. DALL-E 3 [36] addresses this issue by re-captioning

Table 1: Text encoder used for different image generation models.

Methods	Text Encoder	Language
SD1.5 [32]	CLIP ViT-L	English
SD2.0 [32]	OpenCLIP ViT-H	English
DALL-E 2 [30]	CLIP	English
Imagen [34]	T5-XXL	English
MUSE [4]	T5-XXL	English
PixArt-α [5]	T5-XXL	English
LuminaT2X [12]	LLama2-7B	English
Kandinsky-3 [1]	Flan-UL2	Multilingual
SDXL [27]	CLIP ViT-L & OpenCLIP ViT-bigG	English
Playground-v2.5 [18]	CLIP ViT-L & OpenCLIP ViT-bigG	English
eDiff-I [2]	CLIP L & T5-XXL	English
HunyuanDiT [19]	mCLIP & mT5-XL	Multilingual
Stable Diffusion 3 [9]	CLIP ViT-L & OpenCLIP ViT-bigG & T5-XXL	English
Kolors	ChatGLM3-base	Chinese / English

the training dataset using a specialized image captioner. To enhance the prompt-following capability of Kolors, we adopt a similar approach as DALL-E 3, re-captioning the text-image pairs with a multimodal large language model (MLLM).

We propose to evaluate the quality of text descriptions based on the following five criteria:

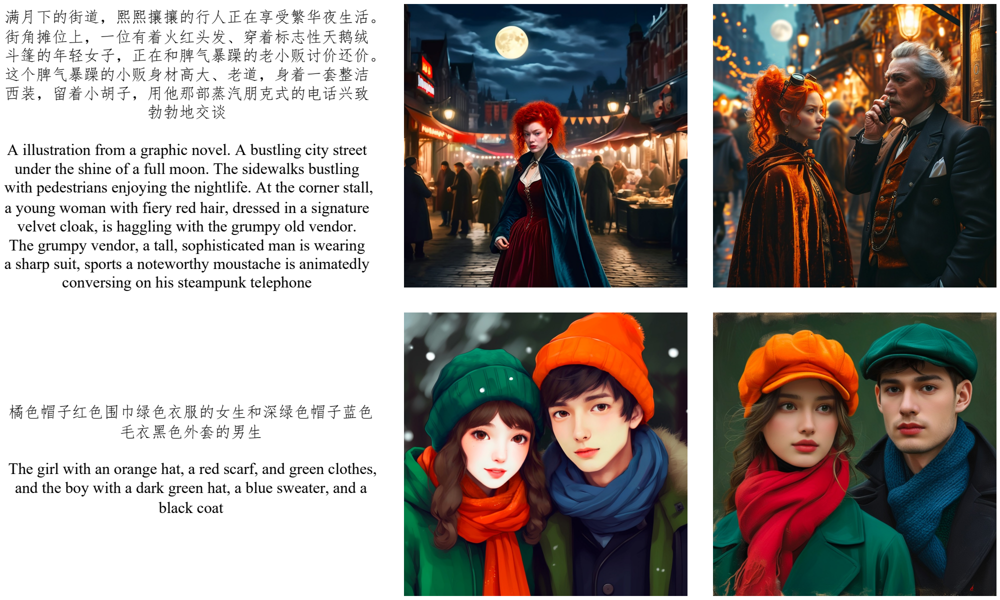
- **Length**: The total number of Chinese characters.
- **Completeness**: The degree to which the text description encompasses the entire image. A score of 5 is assigned if the text describes all objects in the image, whereas a score of 1 is given if it describes less than 30% of the objects.
- **Correlation**: The accuracy of the text description in representing the foreground elements of the image. A score of 5 is awarded if the text describes all foreground objects, while a score of 1 is given if it covers less than 30% of the foreground objects.
- **Hallucination**: The proportion of details or entities mentioned in the text that are not present in the image. A score of 5 indicates no hallucinations in the text, while a score of 1 is assigned if more than 50% of the text is hallucinatory.
- **Subjectivity**: The extent to which the text deviates from describing the visual content of the image and instead conveys subjective impressions. For instance, a sentence like "It gives a feeling of ease and tranquility and makes people feel comfortable" is considered subjective. A score of 5 is given if there is no subjective text, and a score of 1 if more than 50% of the text consists of subjective sentences.
- **Avg**: The average score derived from Completeness, Correlation, Hallucination, and Subjectivity.

We selected five well-known MLLM models and employed ten evaluators to assess 500 images. Notably, LLaVA1.5 [22], CogAgent [14], and CogVLM [40] support Chinese text. However, we found that the generated Chinese captions were inferior to their English counterparts. Consequently, we first generated English image captions and then translated them into Chinese. In contrast, we used the Chinese prompt directly with InternLM-XComposer-7B [49] and GPT-4V.

The captioning performance of the five models is summarized in Table 2. It is evident that GPT-4V achieves the highest performance. However, processing hundreds of millions of images with GPT-4V is prohibitively expensive and time-consuming. Among the four remaining open-source models, LLaVA1.5 and InternLM-XComposer exhibit significantly inferior performance in terms of Completeness and Correlation when compared to CogAgent-VQA and CogVLM-1.1-chat, indicating lower quality of highly detailed descriptions. Additionally, the captions generated by CogVLM-1.1-chat demonstrate fewer instances of Hallucination and Subjectivity. Based on these evaluations, we selected the state-of-the-art vision-language model CogVLM-1.1-chat to generate synthetic detailed

Table 2: The caption performance of different MLLM models.

Methods	Length	Completeness	Correlation	Hallucination	Subjectivity	Avg
LLaVA1.5-13B [22]	33.51	3.90	4.24	4.69	4.21	4.26
InternLM-XComposer-7B [49]	64.71	3.92	4.15	4.69	4.27	4.26
CogAgent-VQA [14]	129.38	4.55	4.50	4.28	4.46	4.45
CogVLM-1.1-chat [40]	99.7	4.54	4.54	4.64	4.51	4.56
GPT4-V	43.89	4.75	4.60	4.78	4.52	4.66


满月下的街道，熙熙攘攘的行人正在享受繁华夜生活。街角摊位上，一位有着火红头发、穿着标志性天鹅绒斗篷的年轻女子，正在和脾气暴躁的老小贩讨价还价。这个脾气暴躁的小贩身材高大、老道，身着一套整洁西装，留着小胡子，用他那部蒸汽朋克式的电话兴致勃勃地交谈
A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. The grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone
橘色帽子红色围巾绿色衣服的女生和深绿色帽子蓝色毛衣黑色外套的男生
The girl with an orange hat, a red scarf, and green clothes, and the boy with a dark green hat, a blue sweater, and a black coat

(a) Kolors with CLIP

(b) Kolors with GLM

Figure 2: Qualitative comparison between (a) Kolors with CLIP and (b) Kolors with GLM.

captions for our extensive training dataset. Considering that MLLMs may fail to identify specific concepts in images that are not present in their knowledge corpus, we employ a strategy of using a 50% original text to 50% synthetic captions ratio. This aligns with the configuration of Stable Diffusion 3.

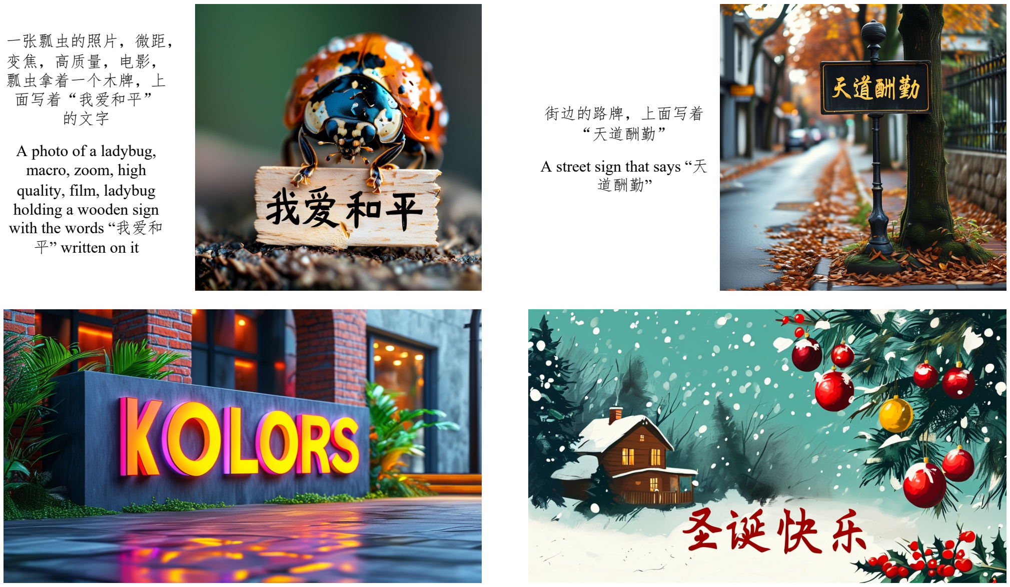
By leveraging fine-grained synthetic captions, Kolors demonstrates a robust ability to follow complex Chinese text. As depicted in Figure 2, we present the results of Kolors with different text encoders on complex prompts. We observe that Kolors, when utilizing GLM, performs well across multiple subjects and detailed attributes. In contrast, Kolors with CLIP fails to generate the vendor and telephone in the top prompt, and there is confusion regarding colors in the bottom prompt.

2.1.3 Enhancement of Chinese Text Rendering Capability.

Text rendering has long been a challenging problem in the field of text-to-image generation. Advanced methods such as DALL-E 3 [3] and Stable Diffusion 3 [9] have demonstrated excellent capabilities in rendering English text. However, current models encounter significant challenges in accurately rendering Chinese text. The underlying reasons for these difficulties are as follows:

1. The extensive set of Chinese characters and the intricate textures of these characters inherently make rendering Chinese text more challenging compared to English.
2. There is a scarcity of sufficient training data that includes Chinese text and related imagery, resulting in inadequate model training and fitting capabilities.

To address this problem, we approach it from two perspectives. Firstly, for the Chinese corpus, we select the 50,000 most frequently used words and construct a training dataset of tens of millions



一张瓢虫的照片，微距，变焦，高质量，电影，瓢虫拿着一个木牌，上面写着“我爱和平”的文字

A photo of a ladybug, macro, zoom, high quality, film, ladybug holding a wooden sign with the words “我爱和平” written on it

街边的路牌，上面写着“天道酬勤”

A street sign that says “天道酬勤”

艺术字招牌设计，绚丽的色彩，高清画质，科技感，写着“KOLORS”

Art signboard design, gorgeous colors, high-definition picture quality, sense of technology, writing “KOLORS”

圣诞贺卡，上面写着“圣诞快乐”

Christmas cards that say “圣诞快乐”

Figure 3: The robust text rendering capabilities of Kolors, particularly in Chinese.

image-text pairs through data synthesis. To ensure effective learning, these synthesized data are incorporated exclusively during the conceptual learning phase. Secondly, to enhance the realism of the generated images, we utilize OCR and multimodal language models to generate new descriptions for real-world images, such as posters and scene texts, resulting in approximately millions samples.

We observe that although the synthetic data in the training dataset initially lacked realism, the realism of the generated text images significantly improves after incorporating real data and high-quality text images into the training process. This enhancement is evident even when some characters appear only in the synthetic data. Additional visualizations are provided in Figure 3.

By systematically addressing the limitations of training data through the integration of both synthetic and real-world data, our approach significantly enhances the quality of Chinese text rendering, thereby paving the way for new advancements in Chinese text image generation.

2.2 Improving Visual Appeal

Although Latent Diffusion Models (LDMs) have been widely adopted by both general users and professional designers, their image generation quality often necessitates additional post-processing steps, such as image upscaling and face restoration. Enhancing the visual quality of LDMs remains a significant challenge. In this work, we address this issue by advancing data and training methodologies.

2.2.1 High Quality Data

The training of Kolors is divided into two distinct phases: the concept learning phase and the quality improvement phase. During the concept learning phase, the model primarily acquires comprehensive knowledge and concepts from a large-scale dataset comprising billions of image-text pairs. The data for this phase is sourced from public datasets (*e.g.* , LAION [35], DataComp [11], JourneyDB [37]) as well as proprietary datasets. By employing a category-balanced strategy, this dataset ensures extensive coverage of a wide range of visual concepts. In the quality improvement phase, the focus shifts to enhancing image details and aesthetics at high resolutions. Previous works [6, 18] have also emphasized the critical importance of data quality during this stage.

To acquire high-quality image-text pairs, we initially apply traditional filters (such as resolution, OCR accuracy, face count, clarity, and aesthetic score) to our dataset, thereby reducing it to approximately tens of millions of images. These images are subsequently subjected to manual annotation, with the annotations categorized into five distinct levels. To mitigate subjective bias, each image is annotated three times, and the final levels are determined through a voting process. The characteristics of the different levels of images are shown as follows:

- **Level 1**: Content deemed unsafe includes images depicting pornography, violence, gore, or horror.
- **Level 2**: Images exhibiting signs of artificial synthesis, such as the presence of logos, watermarks, black or white borders, stitched images, etc.
- **Level 3**: Images with parameter errors, such as blurriness, overexposure, underexposure, or lack of a clear subject.
- **Level 4**: Unremarkable photographs, akin to snapshots taken without much consideration.
- **Level 5**: Photographs of high aesthetic value, which implies that an image should not only possess appropriate exposure, contrast, hue balance, and color saturation, but also convey a sense of narrative.

This methodology ultimately yields millions of Level 5 high-aesthetic images, which are utilized in the quality enhancement phase.

2.2.2 Training on High Resolutions

Diffusion models often underperform at high resolutions due to inadequate disruption of the image during the forward diffusion process. As illustrated in Figure 4, when noise is added following the schedule provided in SDXL [27], low-resolution images degrade into nearly pure noise, whereas high-resolution images tend to retain low-frequency components at the terminal stage. Since the model must start from pure Gaussian noise during inference, this discrepancy can cause inconsistencies between training and inference at high resolutions. Recent studies [20, 15] have proposed methods to address this issue.

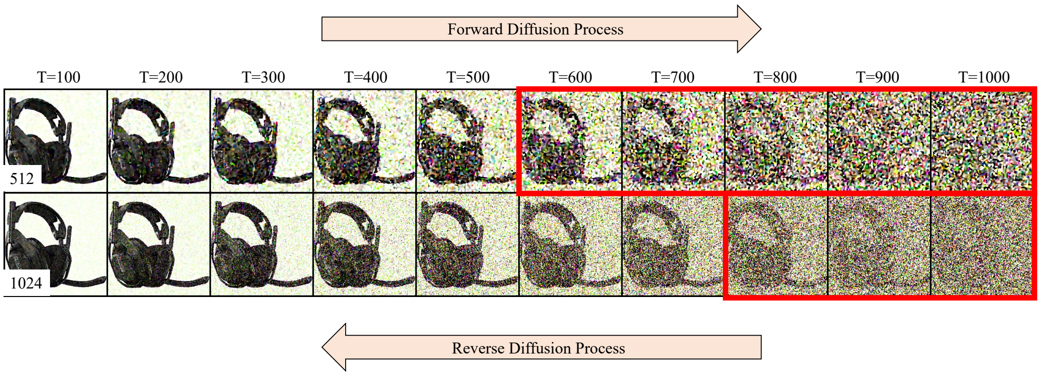


Figure 4: Standard diffusion noise on an image with different resolutions.

In Kolors, we adopt a DDPM-based training approach [13] with an epsilon prediction objective. During the low-resolution training phase for concept learning, we employ the same noise schedule as SDXL [27]. For high-resolution training, we introduce a new schedule, which simply extends the number of steps from the original 1000 to 1100, enabling the model to achieve a lower terminal signal-to-noise ratio. Additionally, we adjust the value of β to maintain the shape of the $\overline{\alpha_t}$ curve, where $\overline{\alpha_t}$ determines $x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon$. As illustrated in Figure 5, our $\overline{\alpha_t}$ trajectory fully encompasses the base schedule's trajectory, while the trajectories of other methods exhibit significant deviations. This indicates that when transitioning from the base schedule used in low resolution, the adaptation and learning difficulty of the new schedule are reduced compared to other schedules.

As illustrated in Figure 6, the quality of generated images has been significantly enhanced by integrating high-quality training data with optimized high-resolution training techniques. Furthermore,

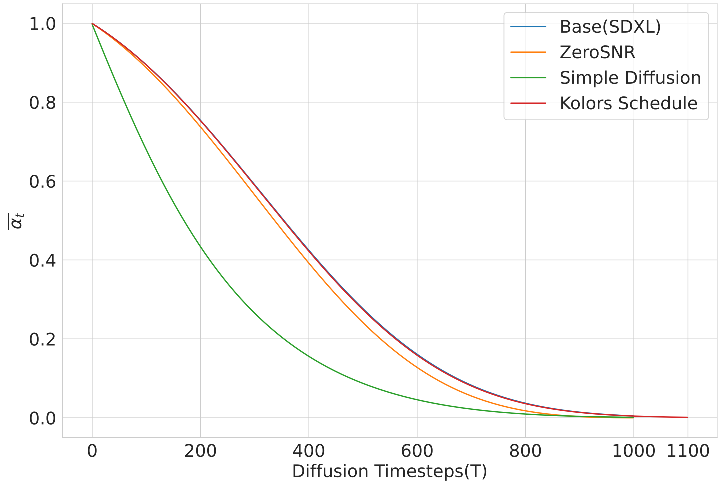


Figure 5: The $\overline{\alpha_t}$ curve of different noise schedules. Noting that the blue trajectory of the base schedule is covered by the red one of our Kolors' schedule.

to enable the model to generate images with diverse aspect ratios, we employ NovelAI's bucketed sampling method [25] during the training process. To conserve training resources, this strategy is exclusively applied during high-resolution training. Examples of images with different resolutions are presented in Figures 1 and 9.

3 Evaluations

To accurately assess the generative capabilities of Kolors, we establish three fundamental evaluation metrics. Firstly, we introduce a new benchmark, *KolorsPrompts*, a prompt set that encompasses multiple categories and distinct challenges. We then conduct a comprehensive evaluation based on human preferences using *KolorsPrompts*. Additionally, we compute two automatic evaluation metrics: (a) the Multi-dimensional Preference Score (MPS) [50] and (b) the well-known image quality assessment metric, FID. We compare Kolors against both open-source models and proprietary systems available on the market.

3.1 *KolorsPrompts*

To comprehensively evaluate text-to-image generation models, we introduce a holistic benchmark *KolorsPrompts*. Specifically, *KolorsPrompts* comprises over a thousand prompts sourced from publicly available datasets, including PartiPrompts [47] and ViLG-300 [10], along with some proprietary prompts. *KolorsPrompts* encompasses 14 prevalent scenarios (*e.g.* , people, food, animals, art, etc.) in the real world. Furthermore, we categorize *KolorsPrompts* into 12 distinct challenges based on the characteristics of the prompts. Each prompt is provided in both Chinese and English versions.

The distribution of *KolorsPrompts* is detailed in Figure 8. The distribution settings of categories and challenges reflect practical usage. The left side of the figure illustrates the category distributions of *KolorsPrompts*, with the category ***People*** accounting for the largest proportion at 29.4%. The right side shows the distribution of challenges, with ***Simple words*** being the most prevalent, accounting for 30.9%.

3.2 Human Evaluation

We provide three human evaluation metrics to assess the model's performance:

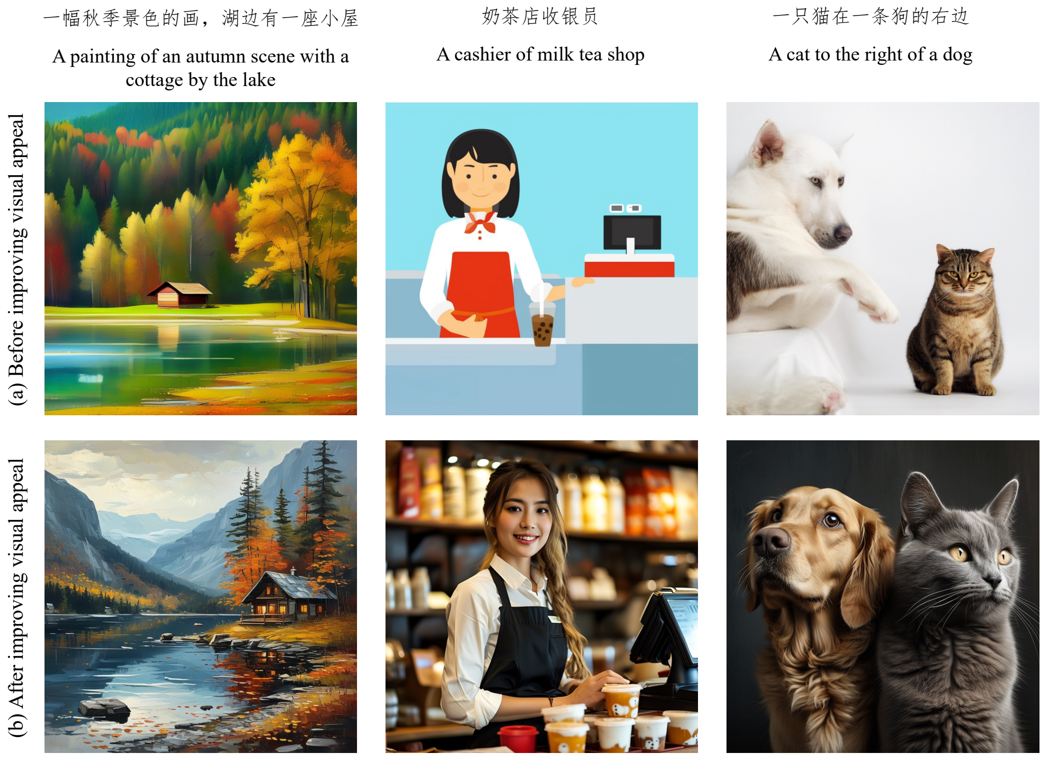

一幅秋季景色的画，湖边有一座小屋
A painting of an autumn scene with a cottage by the lake
奶茶店收银员
A cashier of milk tea shop
一只猫在一条狗的右边
A cat to the right of a dog
(a) Before improving visual appeal
(b) After improving visual appeal

Figure 6: Qualitative comparison between (a) before improving visual appeal and (b) after improving visual appeal.

- **Visual Appeal**. Visual appeal refers to the overall aesthetic quality of a generated image, encompassing various visual elements such as color, shape, texture, and composition to create a pleasing and engaging appearance. In this evaluation, we present users with images generated by different models using the same prompt, without displaying the corresponding text description. This approach allows users to focus exclusively on the visual appeal of the images. Each evaluator rates one image on a scale from 1 to 5, where 5 signifies perfection and 1 denotes the lowest quality.
- **Text Faithfulness**. Text faithfulness measures how accurately a generated image corresponds to its accompanying prompt. Evaluators are instructed to disregard image quality and focus solely on the relevance between the text description and the images. Scoring is conducted on a scale from 1 to 5.
- **Overall Satisfaction**. Overall satisfaction represents a holistic assessment of the image. In this evaluation, the prompt is displayed alongside the image. Evaluators assess the image based on its quality, visual appeal, and the alignment between the prompt and the image, rating it on a scale from 1 to 5.

The model under evaluation generates four images per prompt. We engaged around 50 professional reviewers to assess each image five times according to the specified guidelines. The final score of an image is calculated as the average rating from these five evaluations. Consequently, each image receives three distinct scores for visual appeal, text faithfulness, and overall satisfaction. All images are rendered at a resolution of 1024×1024 pixels.

Considering the high cost of manual evaluation, our human evaluation focuses on the current state-of-the-art text-to-image models, including Adobe Firefly, DALL-E 3 [3], Stable Diffusion 3 [9], Midjourney-v5, Midjourney-v6, and Playground-v2.5 [18]. To ensure optimal performance of the comparison models, we provide the prompts in English for these models, while Kolors receives prompts in Chinese. The detailed results are presented in Figure 8. Kolors achieves the highest overall satisfaction, on par with proprietary models such as Midjourney-v6. Notably, Kolors demonstrates a

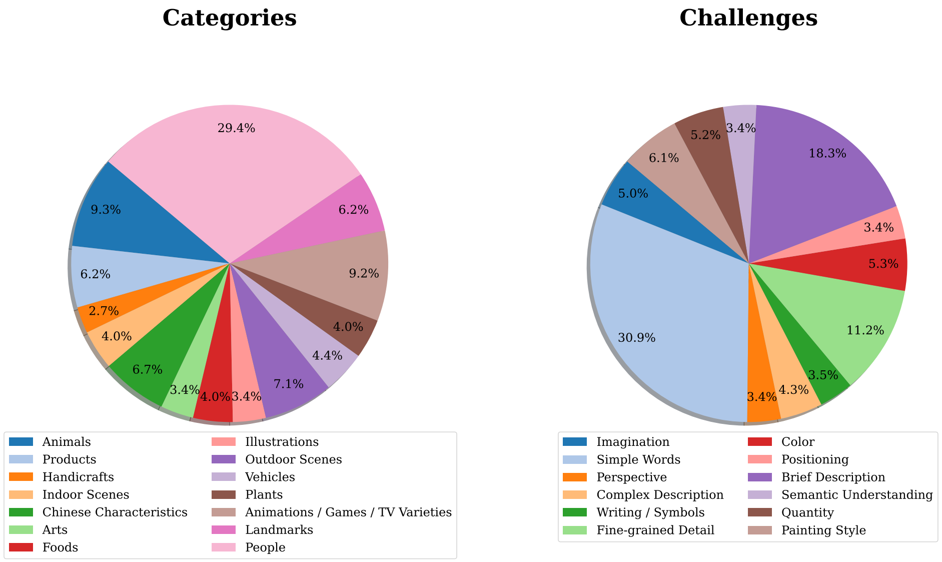


Figure 7: Distribution of *KolorsPrompts*.

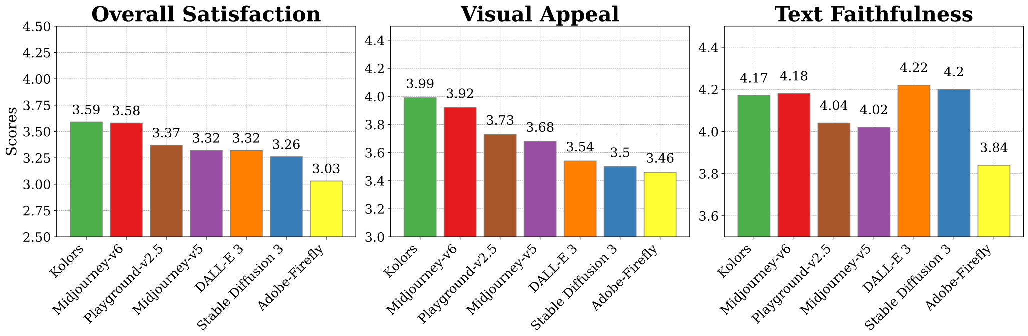


Figure 8: Human evaluations on *KolorsPrompts*.

significant advantage in terms of visual appeal. Additional examples of images generated by Kolors are shown in Figure 9.

3.3 Automatic Evaluation Benchmark

3.3.1 Multi-Dimensional Human Preference Score (MPS)

Current evaluation metrics for text-to-image models predominantly rely on singular measures (*e.g.* , FID, CLIP Score [28]), which inadequately capture human preferences. The Multi-dimensional Human Preference Score (MPS) [50] has been proposed to evaluate text-to-image models from multiple dimensions of human preferences and has demonstrated its effectiveness in text-to-image evaluation. Consequently, we evaluate the above text-to-image models on *KolorsPrompts* benchmark using MPS.

The results of MPS are presented in Table 3. As observed, Kolors achieves the highest performance, consistent with human evaluations. This consistency indicates a strong correlation between human preferences and the MPS scores on the *KolorsPrompts* benchmark.

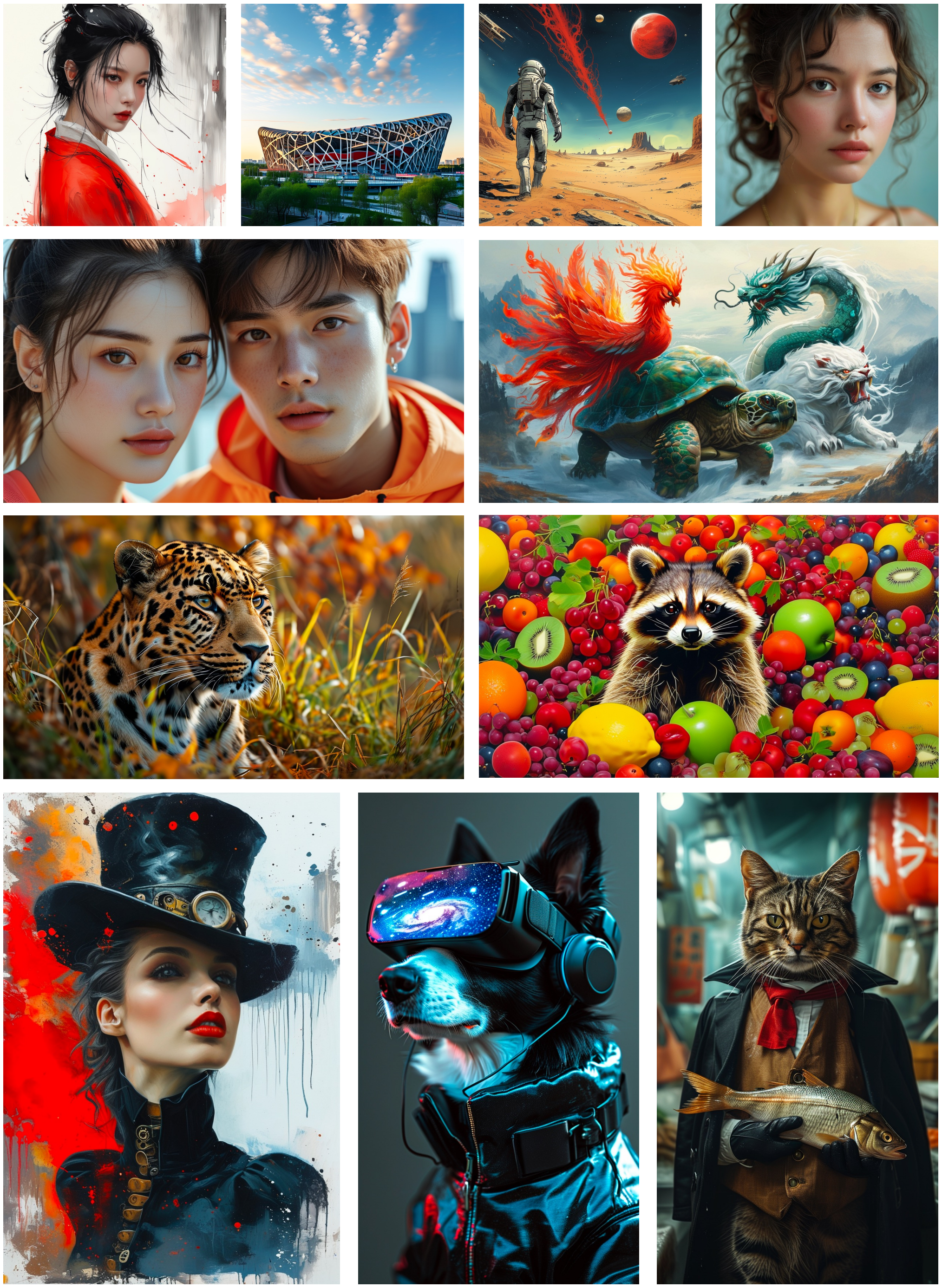


Figure 9: Images generated by Kolors.

Table 3: The overall scores of MPS on *KolorsPrompts*.

Models	Overall MPS↑
Adobe-Firefly	8.5
Stable Diffusion 3 [9]	8.9
DALL-E 3 [3]	9.0
Midjourney-v5	9.4
Playground-v2.5 [18]	9.8
Midjourney-v6	10.2
Kolors	10.3

Table 4: Performance comparison on MS-COCO using zero-shot FID-30K.

Methods	#Params	FID-30K↓
DALL-E [31]	12.0B	27.50
GLIDE [24]	5.0B	12.24
DALL-E 2 [30]	6.5B	10.39
PixArt-α [5]	0.6B	10.65
ParaDiffusion [41]	1.3B	9.64
GigaGAN [16]	0.9B	9.09
SD [32]	0.9B	8.32
Imagen [34]	3.0B	7.27
ERNIE-ViLG 2.0 [10]	22B	6.75
DeepFloyd-IF [7]	4.3B	6.66
RAPHAEL [44]	3.0B	6.61
Kolors	2.6B	23.15

3.3.2 Fidelity Assessment on COCO Dataset

We also evaluate Kolors using the standard evaluation metric for text-to-image generation tasks, namely the Zero-shot FID-30K on the MS-COCO 256×256 [21] validation dataset. Table 4 shows the comparisons between Kolors and other existing models. Kolors achieves slightly higher FID scores, which may not be perceived as a highly competitive result. However, we argue that FID might not be an entirely suitable metric for assessing image quality, as a higher score does not necessarily correlate with superior generated images.

Numerous studies [5, 7, 27, 17, 50] have demonstrated that ***the zero-shot FID on COCO is negatively correlated with visual aesthetics***, and the generative performance of text-to-image models is more accurately assessed by human evaluators than by statistical metrics. These findings highlight the necessity for an automatic evaluation system that aligns with real human preferences, such as MPS [21].

4 Conclusions

In this work, we introduce Kolors, a latent diffusion model built upon the classic U-Net architecture [27]. By leveraging the General Language Model (GLM) and the fine-grained captions generated by CogVLM, Kolors excels in comprehending intricate semantics, particularly those involving multiple entities, and demonstrates superior text rendering capabilities. Moreover, Kolors is trained through two distinct phases: the concept learning phase and the quality improvement phase. By utilizing high-aesthetic data and employing a new schedule for high-resolution image generation, the visual appeal of the resulting high-resolution images is significantly enhanced. Additionally, we propose a new category-balanced benchmark *KolorsPrompts* to comprehensively evaluate text-to-image generation models. Kolors achieves outstanding performance on human evaluations, surpassing the majority of both open-source and proprietary models such as Stable Diffusion 3, Playground-v2.5 and DALL-E 3, and demonstrating performance on par with Midjourney-v6.

We are pleased to announce the public release of the model weights and code of Kolors. In future work, we aim to progressively release various applications and plug-ins for Kolors, including ControlNet [48], IP-Adapter [46], and LCM [23], among others. Additionally, we intend to release a new proprietary diffusion model based on the Transformer architecture [26]. We aim for Kolors to propel the advancement of the text-to-image synthesis community and are dedicated to making substantial contributions to the open-source ecosystem.

References

[1] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report, 2023.

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

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

[4] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023.

[5] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\backslash alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.

[6] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[7] Deepfloyd. Deepfloyd. *URL https://www.deepfloyd.ai/.*, 2023.

[8] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling, 2022.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

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

[11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[12] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, Renrui Zhang, Junlin Xi, Wenqi Shao, Zhengkai Jiang, Tianshuo Yang, Weicai Ye, He Tong, Jingwen He, Yu Qiao, and Hongsheng Li. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers, 2024.

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

[14] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2023.

[15] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

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

[17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.

[18] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[19] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[20] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed, 2024.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

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

[23] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[25] NovelAI. Novelai improvements on stable diffusion. `https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac`, 2022.

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

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[36] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020.

[37] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

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

[39] Chengyu Wang, Zhongjie Duan, Bingyan Liu, Xinyi Zou, Cen Chen, Kui Jia, and Jun Huang. Pai-diffusion: Constructing and serving a family of open chinese diffusion models for text-to-image synthesis on the cloud. *arXiv preprint arXiv:2309.05534*, 2023.

[40] WeihanWang, Wenmeng Yu Qingsong Lv, Yan Wang Wenyi Hong, Ji Qi, Zhuoyi Yang Junhui Ji, Xixuan Song Lei Zhao, Xu Bin Jiazheng Xu, Yuxiao Dong Juanzi Li, and Jie Tang Ming Dingz. Cogvlm: Visual expert for large language models. *arXiv preprint arXiv:5148899*, 2023.

[41] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model, 2023.

[42] Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support. *arXiv preprint arXiv:2401.14688*, 2024.

[43] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

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

[45] Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6648–6656, 2024.

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

[47] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[49] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023.

[50] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation, 2024.