# LDGen: Enhancing Text-to-Image Synthesis via Large Language Model-Driven Language Representation

**Anonymous ACL submission**


Figure 1: Generated image samples from LDGen. We present a composed prompt with each language in a different color, along with the corresponding image that exhibits high aesthetic quality and text-image alignment.

## Abstract

In this paper, we introduce LDGen, a novel method for integrating large language models (LLMs) into existing text-to-image diffusion models while minimizing computational demands. Traditional text encoders, such as CLIP and T5, exhibit limitations in multilingual processing, hindering image generation across diverse languages. We address these challenges by leveraging the advanced capabilities of LLMs. Our approach employs a language representation strategy that applies hierarchical caption optimization and human instruction techniques to derive precise semantic information,. Subsequently, we incorporate a lightweight adapter and a cross-modal refiner to facilitate efficient feature alignment and interaction between LLMs and image features. LDGen reduces training time and enables zero-shot multilingual image generation. Experimental results indicate that our method surpasses baseline models in both prompt adherence and image aesthetic quality, while seamlessly supporting multiple languages.

## 1 Introduction

Text-to-image (T2I) models aim to generate images from text descriptions. (Rombach et al., 2022; Podell et al., 2023; Saharia et al., 2022; Bai et al., 2024; Nichol et al., 2022). Thus, natural language descriptions serve as a critical bridge for conveying user intent and generating visually appealing images that accurately capture the intended semantic information. Despite the impressive performance demonstrated by advanced text-to-image models,

their reliance on text encoders such as CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020), which are primarily tailored for English, constrains their multilingual capabilities due to the linguistic limitations of training datasets.

Recently, large language models (Bai et al., 2023; Liu et al., 2024a; Achiam et al., 2023; GLM et al., 2024; Dubey et al., 2024) have achieved notable success in the field of natural language processing. These models possess advanced language comprehension abilities, enabling them to deeply analyze prompts and provide rich, precise semantic guidance for image generation. Furthermore, many LLMs (Team et al., 2024; Bai et al., 2023; Achiam et al., 2023) are trained on multilingual corpora, granting them the ability to support multiple languages. These advantages have motivated researchers to explore the use of LLMs in text-to-image generation tasks. However, some prior approaches (Xie et al., 2024; Ma et al., 2024) have attempted to directly replace text encoders with LLMs, leading to unstable training processes and significant challenges for researchers with limited computational resources. For instance, ELLA (Hu et al., 2024) and LLM4GEN (Liu et al., 2024c) seek to align LLMs with the CLIP model but require extensive training data to adapt LLMs representations within diffusion models. These methods often treat LLMs features as mere text conditions, thereby failing to fully exploit the comprehensive language understanding capabilities of LLMs. As shown in Appendix A, directly employing LLMs for image descriptions can introduce unintended content, resulting in semantic biases and adversely affecting the output quality of diffusion models.

To effectively address these challenges and integrate large language models into existing text-to-image tasks under resource constraints, we propose LDGen. Our approach enables the efficient incorporation of LLM into current diffusion models based on T5/CLIP text encoders with minimal computational demands. As shown in Fig. 2, we introduce a robust language representation strategy (LRS). By utilizing hierarchical caption optimization and human instruction strategies, LRS fully harnesses the instruction-following, in-context learning, and reasoning capabilities of LLM to accurately derive textual information, thereby enhancing semantic alignment between text and image.

Furthermore, inspired by recent advancements in alignment methods (Hu et al., 2024; Zhao et al., 2024; Tan et al., 2024), we employ a lightweight adapter to align LLM features with T5-XXL, substantially reducing the training time required for text-image alignment. Additionally, we introduce a cross-modal refiner to improve text comprehension and facilitate interaction between LLM and image features. After alignment, the LLM features processed through this refiner exhibit enhanced representational capability. Specifically, the cross-modal refiner integrates self-attention layers, cross-attention layers, and feed-forward neural networks.

By employing this method, LLM can be effectively integrated into existing diffusion models with minimal training. Moreover, the multilingual capabilities of LLM are preserved, enabling zero-shot multilingual image generation without the necessity for training on multilingual text-image datasets. Our experimental results demonstrate that by leveraging the intrinsic features of LLM alongside our innovative modules, LDGen surpasses the prompt comprehension performance of advanced baseline models while seamlessly supporting multiple languages. As shown in Fig. 1, we present several generated images. Our contributions can be summarized as follows:

- We present LDGen, which efficiently integrates LLM into existing text encoder-based diffusion models and supports zero-shot multilingual text-to-image generation.
- We propose a language representation strategy that leverages the capabilities of LLM through hierarchical caption optimization and human instruction strategies.
- We introduce LLM alignment and a cross-modal refiner to achieve LLM feature alignment and enhance interaction between LLM and image features, enhancing the semantic consistency of conditions.

## 2 Related Work

**Text-to-Image.** Recently, denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) have achieved breakthroughs in image synthesis. By mapping the image pixels to a more compact latent space where a denoising network is trained to learn the reverse diffusion process, prominent text-guided generation models have achieved impressive results in terms of image quality and semantic fidelity. Earlier methods (Rombach et al., 2022; Podell et al., 2023) based on the UNet have been tremendously successful in various generative

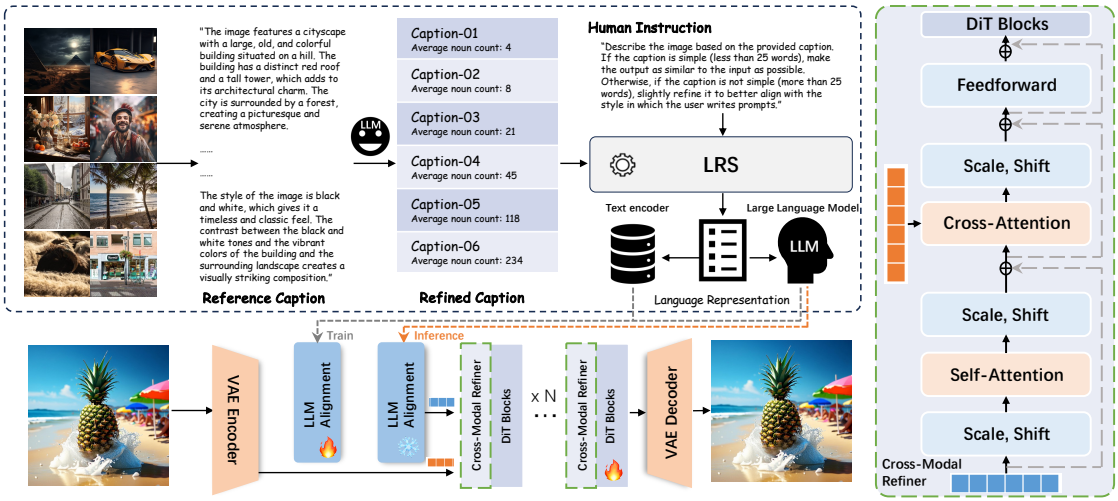



Figure 2: Overview of LDGen. The dashed box shows our language representation strategy, with the bottom is our LLM alignment and cross-modal refiner training process. The detailed design of the cross-modal refiner is shown in the green box on the right.

tasks. With the success of the transformer architecture in various fields, diffusion transformer-based methods (Peebles and Xie, 2023; Gao et al., 2023) are notably developing. Techniques like FLUX (Labs, 2024) and SD3 (Esser et al., 2024) introduced the design of MMBlock to further align text and images during training. PixArt-$\alpha$ (Chen et al., 2023) explored efficient text-to-image training schemes and achieved the first Transformer-based T2I model capable of generating high-quality images at 1024 resolution. Models like Lumina-T2X (Gao et al., 2024) and GenTron (Chen et al., 2024b) extended diffusion transformers from image generation to video generation. Playgroundv3 (PG3) (Liu et al., 2024b) proposed a comprehensive VAE training, caption annotation, and evaluation strategy.

**Large Models in T2I.** The text encoder plays a crucial role in the text-to-image task. In the initial LDM (Rombach et al., 2022), CLIP (Radford et al., 2021) was used as the text encoder, providing the diffusion model with text comprehension capabilities. Later, Imagen (Saharia et al., 2022) discovered that using a large language model with an encoder-only structure like T5 (Raffel et al., 2020) significantly enhanced the model's text understanding. Following this, several works (Chen et al., 2023, 2024a; Sun et al., 2024; Betker et al., 2023; Esser et al., 2024) utilized the T5 series of models as text encoders during pre-training. Additionally, some other works (Liu et al., 2024c; Hu et al., 2024; Zhao et al., 2024; Tan et al., 2024), attempted to adapt the T5 and LLMs (Dubey et al., 2024) to the base models pre-trained based on CLIP. Considering the recent success of decoder-only large language models, some works have sought to apply them in image generation frameworks. PG3 (Liu et al., 2024b) focused on model structure, believing that knowledge in LLMs spans all layers, thus replicating all Transformer blocks from the LLM. LiDiT (Ma et al., 2024), from an application perspective, designed an LLM-infused Diffuser framework to fully exploit the capabilities of LLMs. Sana (Xie et al., 2024), focusing on efficiency, directly used the final layer of LLM features as text encoding features. Kolors (Team, 2024) adapts LLMs for use with SDXL by simply replacing the original CLIP text encoder with ChatGLM. These efforts collectively demonstrate that LLMs still hold significant research potential in the field of image generation.

## 3 Method

### 3.1 Motivation

Text encoding is a pivotal component in text-to-image models, significantly influencing the quality of the generated images. As shown in Fig. 3, the CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020) series models currently dominate the field of text encoders. However, the rapid advancement of large language models (Achiam et al., 2023; Team et al., 2024) is noteworthy. These models employ autoregressive language modeling techniques (Yang, 2019; Black et al., 2022) in unsupervised learning. Through processing vast amounts of

text data, they are beginning to exhibit remarkable reasoning and contextual understanding capabilities. They excel across a range of textual tasks. In particular, LLMs trained on multilingual corpora have demonstrated substantial promise in text-to-image generation tasks. Nonetheless, a critical challenge persists: many existing models rely on CLIP/T5 series text encoders, which are predominantly trained on English corpora and perform effectively. Transitioning to LLMs by replacing the existing text encoders and retraining these models from scratch would involve considerable resource expenditures. To address this issue, we employ LDGen, which seamlessly integrates LLMs into existing diffusion models based on T5/CLIP text encoders, utilizing only a small portion of the initial training resources. These new models not only outperform the originals but also enable zero-shot text-to-image generation across multiple languages.

### 3.2 Language Representation Strategy

Based on the above analysis, while large language models offer substantial advantages, they still encounter several significant challenges. As dialogue models, LLMs employing a decoder-only architecture rely on autoregressive language modeling methods. These models learn linguistic patterns through unsupervised training on large-scale text datasets by predicting the subsequent word in a sequence. However, this characteristic often makes it difficult to control the model outputs, leading to producing a considerable amount of redundant information.

We observe that both LiDiT (Ma et al., 2024) and Sana (Xie et al., 2024) utilize human instructions to help LLMs produce more stable content. However, as shown in Fig. 4, these methods can conflict with the original captions. Incorrect human instructions may cause outputs to deviate from factual accuracy and generate fabricated information, thereby disrupting text-image alignment and potentially decreasing the effectiveness of training.

To address these challenges, we employ a hierarchical captioning strategy. This approach is complemented by extensive human instruction optimization to achieve optimal language representation and enhance semantic alignment between text and images. First, similar to PG3's (Liu et al., 2024b) multi-level image description technique, we utilize the Internvl2-40B model (Chen et al., 2024d,c) to re-caption all image data. We generate six captions of varying lengths, ranging from

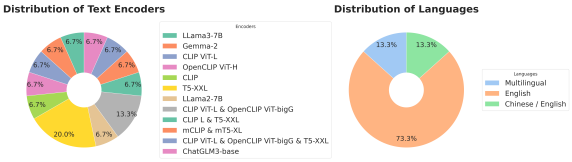



Figure 3: Distribution of text encoder and supported languages. English-based CLIP/T5 series models remain the primary text encoders.


Figure 4: The red words in Sana's generated result highlight elements that do not align with the image. Providing incorrect instructions can change the original caption, potentially creating inaccurate descriptions.

simple to detailed, to comprehensively capture the image content. For detailed captioning prompts, please refer to Appendix Fig. 8, HI-05. During training, these hierarchical captions are randomly sampled and input into the LLM. As shown in Tab. 1, compared to original single-caption methods, LRS enables the model to more effectively capture the hierarchical structure of language concepts while maintaining a high CLIP score.

For these complex and varied-length hierarchical captions, we further refined human instructions to ensure that the LLM's outputs maintain a high CLIP score and avoid generating non-existent information. As shown in Tab. 1, the LLM surprisingly enhances the CLIP scores of the original captions, revealing that our language representation strategy effectively extracts semantic information and enhances text-to-image alignment during model training. To support multilingual text-to-image generation, we evaluated several mainstream LLMs. We selected Qwen (Yang et al., 2024) as our preferred model because it is one of the few trained on multilingual corpora and exhibits exceptional performance in text-related tasks.

### 3.3 LLM Alignment

For pre-trained diffusion models (Chen et al., 2023; Podell et al., 2023), aligning the original text encoder with LLMs features using linear layers is challenging. This is primarily due to the signif-

Table 1: Human Instruction Comparison. Each entry has the CLIP-Score (Hessel et al., 2021) on the left and the LongCLIP-Score (Zhang et al., 2024) on the right, with the average word number is in gray brackets (.). **Original** refers to the initial caption. "HI" indicates outputs from various Human Instruction strategies. The highest scores are highlighted in **bold**, while the second-highest scores are underlined. Scores that surpass the original captions are marked with a gray background.

| | Original | Ours | No-HI | HI-01 | HI-02 | HI-03 | HI-04 | HI-05 |
|---|---|---|---|---|---|---|---|---|
| Caption-1 | 27.65/29.53 | **27.66** / 29.74 | 22.21/27.44 | 22.79/27.95 | 23.33/**30.04** | 22.61/26.59 | 23.06/29.61 | 22.40/27.97 |
| | (4.48) | (7.63) | (204.29) | (335.19) | (87.52) | (171.14) | (254.19) | (224.43) |
| Caption-2 | 29.65/31.49 | **29.66** / 31.63 | 22.89/28.35 | 23.76/30.31 | 24.47/**31.78** | 23.32/28.89 | 24.07/31.25 | 23.41/29.93 |
| | (8.99) | (11.67) | (173.66) | (307.68) | (70.00) | (172.76) | (245.64) | (214.01) |
| Caption-3 | 30.20/33.24 | **29.50/32.92** | 23.75/29.52 | 24.40/31.05 | 25.29/32.77 | 24.31/30.31 | 24.72/32.37 | 24.33/31.03 |
| | (21.71) | (25.25) | (183.91) | (306.33) | (80.78) | (194.68) | (226.49) | (192.64) |
| Caption-4 | 27.53/34.64 | **27.13/33.76** | 24.56/30.03 | 24.67/31.49 | 25.33/33.35 | 24.63/30.42 | 25.01/33.25 | 24.89/32.19 |
| | (45.16) | (46.96) | (249.35) | (329.49) | (116.45) | (253.19) | (205.26) | (167.07) |
| Caption-5 | 25.39/34.43 | **25.40** / **33.74** | 23.87/30.33 | 24.86/31.78 | 25.37/33.37 | 24.38/29.65 | 25.22/33.52 | 25.30/33.20 |
| | (118.06) | (106.18) | (304.45) | (350.39) | (183.10) | (334.34) | (205.27) | (177.34) |
| Caption-6 | 25.42/34.65 | 25.48 /**33.96** | 23.72/31.09 | 25.05/32.18 | 25.36/33.60 | 24.26/30.45 | 25.38/33.89 | **25.59**/33.91 |
| | (118.06) | (106.18) | (304.45) | (350.39) | (183.10) | (334.34) | (205.27) | (177.34) |

icant differences in the output feature spaces of T5/CLIP encoders and LLMs. As a result, directly modifying and training the existing model structure can lead to instability. To address this, we employ a two-step approach: first, we align the feature spaces, then fine-tune the model weights to adapt to the new feature space. This method significantly reduces training time.

Specifically, we first multiply the LLM output by a small coefficient to match the numerical range of T5. This effectively speeds up the feature alignment training. Next, similar to previous methods (Tan et al., 2024), we design a three-layer encoder-decoder Transformer adapter to align the feature spaces of the T5 encoder and LLM output. During the adapter training, we utilize the following alignment loss functions: $\lambda_1 * \mathcal{L}_{\cos} + \lambda_2 * \mathcal{L}_{\text{MSE}}$. The cosine similarity loss aligns the feature space directions, and mean squared error (MSE) loss can further enhance alignment accuracy in terms of numerical range.

By optimizing the alignment loss, we achieve a rough alignment between LLM and T5 output feature spaces. This allows us to quickly integrate the LLM into the pre-trained diffusion model, enhancing its overall performance and adaptability.

### 3.4 Cross-Modal Refiner

To improve text comprehension and facilitate interaction between LLM features and image features, we introduce a lightweight module called the cross-modal refiner. This module employs a sequence of components to optimize and refine LLM feature representations, enabling efficient integration of text and image features. As shown in Fig. 2, it includes elements such as self-attention mechanisms, cross-attention mechanisms, feedforward neural networks, residual connections, normalization layers, and learnable scaling factors.

To enhance the interaction between image and text features, the cross-attention layer serves as a pivotal component of modal interaction. This layer utilizes LLM features as queries, with latent image features acting as keys and values, to facilitate deep interaction between text and image elements. This design enables the refinement and adjustment of text features based on relevant image information, thereby enhancing the model's understanding of cross-modal content. Learnable scaling factors allow the model to gradually balance between original and optimized features during training, ensuring a seamless transition from pre-trained weights to new LLM input features. This mechanism effectively integrates the original LLM's robust semantic understanding into the pre-trained models, boosting overall performance.

The cross-modal refiner module preserves the original LLM features and effectively integrates image-related information to produce richer, semantically aligned conditional representations. This approach allows us to efficiently integrate the LLM into existing diffusion models within relatively short training times, providing highly semantically aligned conditional information for text-to-image generation tasks, significantly enhancing the quality and relevance of generated results.

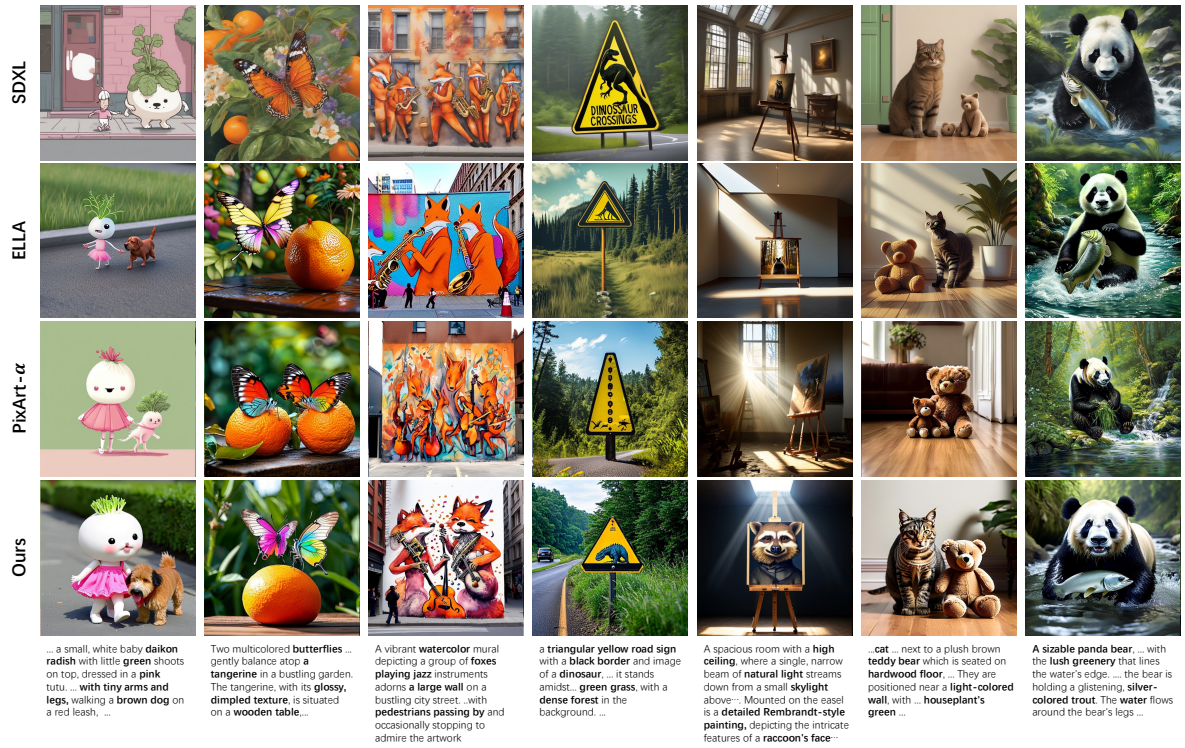



Figure 5: Comparison of our method with recent enhancement generative models ELLA (Hu et al., 2024), baseline Models SDXL (Podell et al., 2023) and PixArt-$\alpha$ (Chen et al., 2023). Our method achieves the best results in terms of instruction adherence and visual appeal.

Table 2: Quantitative comparison results on DPG-Bench. Note that we support multiple languages.

| Method | Param | Multi-Ling | DPG-Bench |
|---|---|---|---|
| SD1.5 (Rombach et al., 2022) | $0.86B$ | × | 61.18 |
| SDv2.1 (Rombach et al., 2022) | $0.89B$ | × | 68.09 |
| LlamaGen (Sun et al., 2024) | $0.78B$ | × | 65.16 |
| HART (Tang et al., 2024) | $0.73B$ | × | 80.89 |
| Sana (Xie et al., 2024) | $0.60B$ | ✓ | 83.6 |
| ELLA (Hu et al., 2024) | $0.93B$ | × | 80.79 |
| LLM4GEN(Liu et al., 2024c) | $0.86B$ | × | 67.34 |
| Pixart-$\alpha$ (Chen et al., 2023) | $0.61B$ | × | 71.11 |
| Ours | $0.63B$ | ✓ | 80.57 |
| SD3-Medium (Esser et al., 2024) | $2.0B$ | × | 84.08 |
| SDXL (Podell et al., 2023) | $2.6B$ | × | 74.65 |
| Janus (Wu et al., 2024) | $1.3B$ | × | 79.68 |
| Janus-Pro (Chen et al., 2025) | $7B$ | × | 84.19 |
| Emu-3 (Wang et al., 2024) | $8.0B$ | × | 80.60 |
| DALL-E 3 (Betker et al., 2023) | – | × | 83.50 |
| FLUX-Dev (Labs, 2024) | $12.0B$ | × | 84.0 |

## 4 Experiments

**Model Details.** Our method is based on the work of PixArt-$\alpha$ (Chen et al., 2023), which is a classic diffusion transformer text-to-image model. It uses the T5-XXL text encoder (Raffel et al., 2020) and has demonstrated excellent performance. We use Qwen2.5-7B-Instruct (Yang et al., 2024) as the LLM and adopt the output features from the last layer, which has a dimension of 3584. The VAE remains consistent with PixArt-$\alpha$ (Chen et al., 2023). For the LLM feature alignment module, we employ a 3-layer encoder-decoder transformer structure, which includes linear layers to align the LLM dimension of 3584 with the T5 dimension of 4096. The cross-modal refiner uses only one block.

**Training Details.** To reduce computational resources, we've structured our training process into several key stages. First, we train the LLM feature alignment module using approximately 80 million text entries from internal image descriptions, with about 20% of this data being multilingual. Given that T5-XXL (Raffel et al., 2020) doesn't support multiple languages, we align the multilingual features from the LLM output with the English output features of T5-XXL (Raffel et al., 2020). This initial phase consumes around 80 A100 GPU days. Next, drawing inspiration from PixArt-$\alpha$'s training methodology, we adapt our model to 512 resolution and fine-tune it using 24 million text-image pairs. To minimize dataset-specific biases in training, we maintain a data scale similar to PixArt-$\alpha$'s (Chen et al., 2023) original approach and incorporate various datasets with overlapping ranges, such as JourneyDB (Sun et al., 2023). In the final stage, we continue training at a 1024 resolution, utilizing 14


Figure 6: Multilingual qualitative visualization results. For each panel's eight images, we generate them using eight different languages but only display the prompt in one of the languages used. Note that LDGen uses only English prompts during training but achieves zero-shot multilingual generation due to the capabilities of the LLM.

Table 3: We compare our method with baseline methods and fine-tuned baseline methods on DPG-Bench and Geneval, demonstrating the effectiveness of our approach.

| Method | Param | DPG-Bench (Hu et al., 2024) | | | | | Geneval(Ghosh et al., 2023) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Global | Entity | Attri. | Other | Overall | Single Obj. | Two Obj. | Counting | Color Attri. | Overall |
| Pixart-$\alpha$ (Chen et al., 2023) | 0.61B | 74.97 | 79.32 | 78.60 | 76.69 | 71.11 | **0.98** | 0.50 | **0.44** | 0.07 | 0.48 |
| Pixart-$\alpha$(fine-tuned) | 0.61B | 83.18 | 84.06 | 84.07 | 83.61 | 75.05 | 0.95 | 0.37 | 0.37 | 0.43 | 0.46 |
| Ours | 0.63B | **85.88** | **87.83** | **85.21** | **87.85** | **80.57** | 0.88 | **0.55** | 0.35 | 0.42 | **0.51** |

million aesthetic data entries. The entire training process requires approximately 120 A100 GPU days. The count of GPU days excludes the time for T5, Qwen, and VAE feature extraction. LDGen takes only approximately 26% of the GPU days compared to PixArt-$\alpha$.

**Evaluation Metrics.** We evaluate our approach using two publicly available benchmarks: Geneval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024). Geneval is a challenging text-to-image generation benchmark designed to showcase a model's comprehensive generative capabilities through detailed instance-level analysis. DPG-Bench, comprises 1,065 semantically dense long prompts, aimed at evaluating model performance in complex semantic alignment. These two datasets provide a comprehensive assessment of generative models from different perspectives.

### 4.1 Performance Comparison and Analysis

We focus on evaluating the performance of our method compared to the baseline model, PixArt-$\alpha$ (Chen et al., 2023). As shown in Tab. 2 and Tab. 3, we utilize two evaluation benchmarks, DPG-Bench (Hu et al., 2024) and Geneval (Ghosh et al., 2023), to thoroughly assess image-text consistency.

Furthermore, we compare our results with advanced models such as the Stable Diffusion series and enhancement methods like ELLA (Hu et al., 2024) and LLM4GEN (Liu et al., 2024c). Our model not only surpasses these baseline models but also achieves approximately a 13% performance improvement on DPG-Bench compared to PixArt-$\alpha$, approaching the metrics of some larger-scale models. For the Geneval results, we notice that while single-object scores might decrease due to the LLM's data alignment scale being significantly

smaller than the hundreds of millions of samples used for text encoder training, we see significant improvements in multiple aspects, such as color attributes, with the LLM's use.

Although we have made progress, there remains a gap when compared to state-of-the-art models such as HART (Tang et al., 2024) and Sana (Xie et al., 2024), which are trained from scratch with extensive resources and incorporate cutting-edge techniques. Nevertheless, our method achieves significant performance gains on the base model with relatively minimal overhead. Tab. 2 presents our evaluation scores across different languages. Even without using multilingual image-text pairs during training, our model achieves a score of 61.3 in some common languages, nearly matching the 61.2 of certain English-trained image generation models (like SD1.5 (Rombach et al., 2022)). As shown in Tab. 4, we conduct a multilingual generation comparison with Sana and additionally support languages that are not supported by Sana.

As shown in Fig. 5, we present visual comparisons with other enhancement methods like ELLA (Hu et al., 2024) and LLM4GEN (Liu et al., 2024c), as well as the baseline PixArt-$\alpha$. Our method exhibits significant improvements in both aesthetics and text alignment, attributed to the integration of an LLM model with robust comprehension capabilities. Even without employing multilingual image-text data during fine-tuning, our model can generate aesthetically pleasing, instruction-following images in multiple languages.

As shown in Fig. 6, we present generation results in eight languages, displayed from top left to bottom right: German, Spanish, Portuguese, Russian, Italian, Korean, English, and Arabic. Although the model may not generate high-fidelity details across different languages, it is still capable of creating many common scenes and objects.

### 4.2 Ablation Study

In this section, we validate our language representation strategy, LLM alignment module, and cross-modal refiner. First, we conduct a detailed ablation analysis of our Human Instruction (HI) design, with specific details provided in the appendix. Some captions' length exceed CLIP's evaluation capacity, but with LongCLIP (Zhang et al., 2024) supporting up to 248 tokens, we use the LongCLIP score as an additional metric. We randomly select 5,000 samples from the training dataset for calculating their CLIPScore (Hessel et al., 2021) and LongCLIP-Score. As shown in Tab. 1, our HI strategy significantly enhances the CLIP scores of the original captions, demonstrating that our language representation strategy accurately extracts text embeddings and effectively improves text-image alignment during model training.

Table 4: Quantitative comparisons of multilingual generation results. We additionally support some languages that are not supported by Sana.

| Language | Overall↑ | Glob. | Enti. | Attr. | Rela. | Other |
|---|---|---|---|---|---|---|
| Korean (Sana) | 10.6 | 20.3 | 21.3 | 20.1 | 20.5 | 23.7 |
| Korean (Ours) | **50.5** | **73.8** | **63.6** | **68.1** | **70.4** | **68.6** |
| Arabic (Sana) | 12.5 | 22.1 | 26.1 | 23.8 | 25.4 | 31.2 |
| Arabic (Ours) | **50.0** | **64.4** | **66.3** | **66.4** | **72.9** | **66.5** |
| Russian (Sana) | 42.2 | 57.5 | 57.2 | 56.6 | 59.7 | 62.2 |
| Russian (Ours) | **55.9** | **76.1** | **70.8** | **71.4** | **73.5** | **70.2** |
| Spanish (Sana) | **67.4** | **78.9** | **78.1** | **79.6** | 79.8 | 75.3 |
| Spanish (Ours) | 61.3 | 74.1 | 72.0 | 76.7 | **80.3** | **77.9** |

Although our training data size is similar to PixArt-$\alpha$, to eliminate the potential benefits of extra data, we fine-tune the original PixArt-$\alpha$ weights using the T5-XXL (Raffel et al., 2020) with the same training data. As shown in Tab. 3, our method remains superior to this fine-tuned model, validating the effectiveness of our LLM alignment module and cross-modal refiner.

## 5 Conclusion

This paper presents LDGen, which integrates LLMs with diffusion models to enhance text-to-image generation. By using the language representation strategy, LLM alignment module, and cross-modal refiner, we improve semantic alignment between text and images, reduce training demands, and enable zero-shot multilingual generation. Experiments indicate the superiority of LDGen and provide new insights into LLM-T2I tasks.

## 6 Limitations

Our work integrates LLM into diffusion models with text encoders, enhancing text-image alignment and enabling excellent zero-shot multilingual image generation using limited resources. However, our LLM alignment training data is smaller compared to classic text encoders, potentially affecting the understanding of complex prompts and alignment for certain concepts. Additionally, uneven multilingual corpora distribution leads to varied performance across languages. We plan to expand training data in the future to address these issues.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. 2024. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *arXiv preprint arXiv:2410.08261*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024a. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. 2024b. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. 2024. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *Preprint*, arXiv:2403.05135.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. 2024b. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*.

Mushui Liu, Yuhang Ma, Xinfeng Zhang, Yang Zhen, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. 2024c. Llm4gen: Leveraging semantic representation of llms for text-to-image generation.

Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. 2024. Exploring the role of large language models in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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

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

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

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. 2023. Journeydb: a benchmark for generative image understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49659–49678.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Zhiyu Tan, Mengping Yang, Luozheng Qin, Hao Yang, Ye Qian, Qiang Zhou, Cheng Zhang, and Hao Li. 2024. An empirical study and analysis of text-to-image generation using large language model-powered textual representation. In *European Conference on Computer Vision*, pages 472–489. Springer.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. 2024. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Kolors Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. 2024. Bridging different language models and generative vision models for text-to-image generation. In *European Conference on Computer Vision*, pages 70–86. Springer.

# A Appendix

In the appendix, we provide a more comprehensive analysis of the main text, enriched with additional details to enhance understanding.

In Fig. 7, we provide a detailed comparison between our method and the baseline method, Pixart-$\alpha$ (Chen et al., 2023), which demonstrates weaker text comprehension capabilities. Our approach shows significant improvements in terms of aesthetic quality and prompt adherence.

In Fig 8, we perform an extensive comparison utilizing the Human Instruction with the Qwen2.5-7B-Instruct (Yang et al., 2024) of a large language model (LLM), consistent with the version applied in the main text. Our Human Instruction method ensures that the LLM's outputs not only sustain a high CLIP score but also avoid generating non-existent information. Furthermore, this method enhances the accuracy of text embeddings, leading to more reliable outcomes.

Fig. 9 displays more multilingual generation results. Although some images show slight deficiencies in adhering to the prompts, they still produce outstanding results for many common scene.

Fig. 10 showcases images produced from multiple perspectives, including color, theme, style, etc. These varied perspectives effectively illustrate the effectiveness and adaptability of LDGen,


Figure 7: More comparisons with Pixart-$\alpha$. Our method achieves better results in terms of prompt adherence and visua appeal.

**Original Caption**

**Input Caption:** A close-up photograph of a mother monkey and her two young offspring, likely Rhesus macaques, sitting on a rock with a lush green background. The mother monkey, positioned on the left, has a light brown fur coat with a slightly darker shade on her back and a lighter tone on her belly. Her fur appears smooth and well-groomed, and her body is angled slightly towards the camera, showing her profile. She has a calm and protective demeanor, with one of her arms extended to support the first baby monkey, which is sitting on the rock in front of her. The baby monkey, with a mix of brown and black fur, has large, expressive eyes and a curious expression, looking directly at the camera. Its fur is shorter and spikier, especially on the head, and it has a small, pink nose and mouth. The second baby monkey, slightly behind the first, is also looking at the camera, with its head turned to the side, showing a similar mix of brown and black fur and a playful, inquisitive look. The baby monkey's fur is also short and spiky, and it is nestled close to the mother, with one arm resting on the rock and the other on the mother's arm. The background is out of focus, featuring a natural, green environment with hints of sunlight filtering through the leaves, creating a soft, blurred effect that highlights the subjects in the foreground. The lighting is natural and bright, casting gentle shadows and giving the image a warm, serene atmosphere. The composition is centered, with the subjects occupying the lower half of the frame, emphasizing their connection and the nurturing relationship between the mother and her young."

**No Human Instruction**

**Output Caption:** This description paints a vivid and heartwarming scene of a mother Rhesus macaque and her two young offspring. The detailed observation of the monkeys' physical features and expressions adds depth to the imagery, making it almost lifelike.
Here's a summary of the key elements:
1. Setting: A close-up of a mother monkey and her two young offspring on a rock, with a lush green background.
2. Mother Monkey:
   - Light brown fur coat with a darker back and lighter belly.
   - Smooth, well-groomed fur.
   - Calm and protective demeanor.
   - Body angled slightly towards the camera, showing her profile.
   - Supporting the first baby monkey with an extended arm.
3. First Baby Monkey:
   - Mix of brown and black fur.
   - Large, expressive eyes and a curious expression.
   - Short, spiky fur, especially on the head.
   - Small, pink nose and mouth.
4. Second Baby Monkey:
   - Similar mix of brown and black fur.
   - Playful, inquisitive look.
   - Head turned to the side.
   - Short, spiky fur.
   - Nestled close to the mother, with one arm resting on the rock and the other on the mother's arm.
5. Background:
   - Out of focus, featuring a natural, green environment.
   - Hints of sunlight filtering through the leaves, creating a soft, blurred effect.
6. Lighting:
   - Natural and bright.
   - Gentle shadows.
   - Warm, serene atmosphere.
7. Composition:
   - Centered, with the subjects occupying the lower half of the frame.
   - Emphasizes the connection and nurturing relationship between the mother and her young.

**Human Instruction-01**

Describe the image by detailing the color, shape, size, texture, quantity, text, and spatial relationships of the objects.

**Output Caption:** In this close-up photograph, the scene captures the tender interaction between a mother monkey and her two young offspring, likely Rhesus macaques, set against a lush green backdrop.

Color:
- Mother Monkey: Her fur is predominantly light brown, with a slightly darker shade on her back and a lighter tone on her belly. The background features a rich, vibrant green, with hints of sunlight filtering through, adding a soft, warm glow.
- Baby Monkeys: Both have a mix of brown and black fur, with the younger one having slightly more pronounced darker patches. Their fur is described as shorter and spikier, especially on the head, with a small, pink nose and mouth.

Shape:
- Mother Monkey: She is positioned on the left, with a smooth, well-groomed fur coat. Her body is slightly angled towards the camera, showing her profile, and she has a calm and protective demeanor.
- Baby Monkeys: They are both seated on the rock, with one (the first) sitting in front of the mother and the other (the second) behind the first, nestled close to the mother. Their heads are turned, showing a curious and playful expression.

Size:
- The mother monkey is larger than the baby monkeys, with the babies being smaller and closer in size to each other. The mother occupies a significant portion of the lower half of the frame, while the babies are positioned in front of her.

Texture:
- The mother's fur is described as smooth and well-groomed, suggesting a clean and healthy appearance. The babies' fur is shorter and spikier, indicating a younger age and perhaps a different grooming process.

Quantity:
- There are three monkeys in the image: one mother and two baby monkeys.

**Human Instruction-02**

Describe the image based on the provided caption. If the caption is simple (less than 25 words), make the output as similar to the input as possible. Otherwise, if the caption is not simple (more than 25 words), slightly refine it to better align with the style in which the user writes prompts.

**Human Instruction-03**

Given a user prompt, generate an "Enhanced prompt" that provides detailed visual descriptions suitable for image generation. Evaluate the level of detail in the user prompt. If the prompt is simple, focus on adding specifics about colors, shapes, sizes, textures, and spatial relationships to create vivid and concrete scenes. The user prompt:

**Human Instruction-04**

Describe the image by detailing the color, general, image type, text, position, relation, relative position, entity, entity size, entity shape, count, emotion, blur, image artifacts, proper noun (world knowledge), color palette, and color grading. Present this description in a complete paragraph.

**Human Instruction-05**

As an image description expert, generate a detailed description of the image, with the first sentence including all the noteworthy details of the image. The following detailed content description focuses on describing the content details of the image, and should be extremely specific and detailed. Make sure the description can reflect the following image elements:
Image style type (such as scene, photo, portrait, object close-up, promotional image, photography, animation scene, illustration, black and white stills, 3D art, indoor, outdoor, etc.). Detailed subject description, as specific as possible, including their name, shape, quantity, specific color, behavior, relative size, position relationship (such as left and right, up and down, front and back, etc.), object material (such as metal, wood, tile, etc.), texture (such as smooth, rough, wrinkled, creased, cracked, etc.) and lighting details (such as green neon, etc.). For human subjects, describe their facial details, accessories, clothing and other necessary information. For other subjects, have relevant world background knowledge, such as proper nouns (The Bund), names of artworks (Mona Lisa), TV shows (Star Wars), etc. The background and its relationship to the subject. Style information (such as animation, 3D, digital art, surrealism, digital, etc.). Composition and shooting information (such as low angle, wide angle, medium shot, long shot, bird's eye view, fisheye lens, top to bottom, looking up, etc.).

When constructing the description, strictly follow the following five principles:
1. Directly output extremely detailed and complete content, including all necessary information. The expression is smooth and natural, like a human describing the image.
2. Don't use "the image" or similar words.
3. No paragraph titles, tight descriptions, do not separate descriptions, do not mark paragraphs, no more than six paragraphs, but must be detailed and specific.
4. World knowledge needs to be clear before description. If it is not clear, do not describe it.
5. Text information, complete and detailed description.

**Human Instruction-Ours**

"Describe the image based on the provided caption. If the caption is simple (less than 25 words), make the output as similar to the input as possible. Otherwise, if the caption is not simple (more than 25 words), slightly refine it to better align with the style in which the user writes prompts."

Figure 8: We provide detailed comparisons using human instructions ranging from simple to complex, comprehensively evaluating the effectiveness of our method.

**Arabic:** ثمة طائر اسود ملفت للنظر له ريش لامع يجلس فوق بتلات برتقالية نابضة بالحياة لزهرة طائر الجنة. وتتمركز هذه الزهرة الفريدة وسط مناظر صحراوية قاحلة، حيث تنتشر مختلف انواع الصبار والنباتات المتناثرة في الاراضي الرملية. في الخلفية، تُطلق الشمس وهجا دافئا على الكثبان الرملية المتموجة البعيدة.

**Korean:** 널찍한 포장지 현관이 매달린 꽃바구니로 장식된 매력적인 백색의 시골집을 묘사한 그림 같은 그림. 집은 푸르른 녹지를 배경으로 자갈이 깔린 오솔길이 반겨주는 앞 계단으로 이어진다. 현관 난간은 복잡하게 설계되었고, 집의 창문은 전통적인 셔터를 자랑해 진기한의 미학을 더했다.

**Russian:** Современный номер с элегантным монитором с плоским экраном расположен прямо на низком деревянном столе, рядом с пышным серым диваном. Рядом с диваном, есть соответствующий стул, создавая уютный гостиный уголок. Стол, расположенный между стулом и диваном, поддерживает монитор и также находится в пределах досягаемости для сидящих. Над монитором висит современная лампа, дающая достаточно света и добавляющая изящность в установку. Рядом со стулом, диван представляет собой идеальное место для отдыха, сохраняя пространственную гармонию с остальной мебелью.

**German:** Eine höchst komplexe und dynamische städtelandschaft spiegelt eine mischung aus moebius' fantasischem design und dem genauen animalischen stil des neuen meeres wider. Neonfarben leuchteten auf den straßen und spiegelten sich hinter dem Regen auf dem glatten weg. Die hohen wolkenkratzern und glänzendes fenster brechen in den sternenhimmel ein, denn die kunst gab auch weiterhin großartige aufmerksamkeit und lob.

**French:** Deux femmes élégamment habillées, ornées de robes Renaissance complexes aux manches bouffantes et riches broderies, tiennent un smartphone élégant et moderne pour capturer un selfie. Leur tenue comporte des teintes profondes de rouge et d'or, contrastant avec le lustre métallique du téléphone. Elles se trouvent dans une pièce aux éléments architecturaux classiques, dont une grande fenêtre qui les baignera de lumière naturelle.

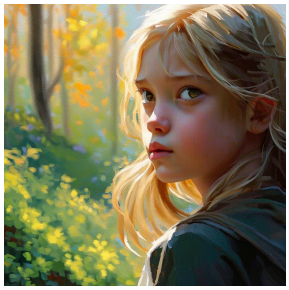
**Italian:** Una bella scatola da regalo avvolta in carta d'argento scintillante e fissata con un nastro rosso brillante, posizionata a sinistra di un lussuoso albero di natale verde ornato da luci scintillanti e ornamenti dorati. L'albero si trova su una gonna morbida bianca che contrasta con il pavimento di legno scuro, e sparpagliata intorno sono regali più piccoli che si aggiungono alla scena festiva.

**English:** A vibrant red bus is parked on a bustling city street, its size overshadowing the adjacent white van. The street is lined with a mix of small shops and residential buildings, each with their own unique facades. The vehicles are positioned parallel to the curb, with the bus's destination sign clearly visible above its windshield.

**Spanish:** Una mujer mayor con el pelo gris rechasta los hombros y gafas redondas de color metal se sienta cómodamente en un sillde felpa. Ella lleva una chaquede de lavansobre una blublanca, y un collar de plata se puede ver alrededor de su cuello. En su regazo descansa un libro abierto de tapa dura, y a su lado, una pequeña mesa de madera sostiene una taza de té y un platillo de cerámica.

Figure 9: More multilingual qualitative visualization results. For each panel's eight images, we generate them using eight different languages but only display the prompt in one of the languages used.

A whimsical scene at the beach where a pineapple, complete with its spiky green leaves, is balanced atop a vibrant blue wave as if it were surfing. The pineapple's textured, golden-brown skin glistens with droplets of ocean water. In the background, the sandy shore is dotted with colorful beach umbrellas and sunbathers enjoying the sunny day.",



In the art piece, a realistically depicted young girl with flowing blonde hair gazes intently into the distance, her eyes reflecting the vibrant hues of a spring forest. The verdant greens and soft pastels of the budding trees are captured in subtle brushstrokes, giving the scene a serene and tranquil atmosphere. The minimalist composition focuses on the girl's expression of wonder and the lush woodland background, while the texture of the oil paint adds depth and richness to the canvas.

A detailed photograph captures the intricate features of a pharaoh statue adorned with unconventional accessories. The statue is wearing steampunk glasses that have intricate bronze gears and round, reflective lenses. It is also dressed in a stark white t-shirt that contrasts with a dark, textured leather jacket draped over its shoulders …… The background is a simple, unobtrusive blur, drawing all attention to the anachronistic ensemble of the pharaoh.

A digital illustration of a girl features her with vibrant rainbow-colored hair that cascades smoothly down her shoulders. She has two spiraling unicorn horns emerging from her forehead, adding a fantastical element to the portrait. Fresh, vivid colors blend seamlessly in gradients across the composition, showcasing a high level of detail and skill. The image is in sharp focus, with rim lighting that highlights the contours of her face and creates a sense of depth against the softly blurred background.

A young woman with freckles wearing a straw hat, standing in a golden wheat field.

car made out of vegetables.

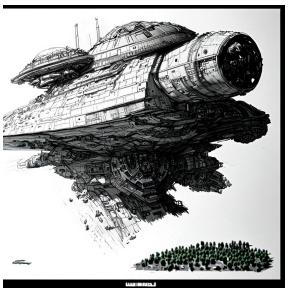
Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

dreamlike digital art captures a vibrant, kaleidoscopic lion in a lush rainforest.

4k dslr image of a lemur wearing a red magician hat and a blue coat performing magic tricks with cards in a garden.

A vibrant yellow banana-shaped couch sits in a cozy living room, its curve cradling a pile of colorful cushions. on the wooden floor, a patterned rug adds a touch of eclectic charm, and a potted plant sits in the corner, reaching towards the sunlight filtering through the window.

A striking an origami pig sits atop the vibrant orange petals of a Bird of Paradise flower. The unique flower is positioned in the midst of an arid desert landscape, with various cacti and sparse vegetation dotting the sandy ground. In the background, the sun casts a warm glow on the distant rolling dunes.



A photo of detailed pen and ink drawing of a massive complex alien space ship above a farm in the middle of nowhere

In the renowned portrait, the subject, known as the Girl with a Pearl Earring, is actually adorned with a pearl drop earring rather than a golden hoop. The soft texture of her pale skin contrasts with the dark, liquid-like background, while her blue and gold turban adds a touch of vibrant color to the composition. Light gently caresses her face, highlighting the luminescent pearl that gracefully hangs from her earlobe.

A focused individual with a blue denim jacket is strumming an electric guitar amidst the quietude of a library. Surrounded by towering wooden bookshelves filled with an array of books, he is seated on a simple chair with a burgundy cushion. His guitar, a sleek black instrument with silvery strings, catches the light from the overhead lamps as he creates a melody in this uncommon setting.

A surreal image capturing an astronaut in a white space suit, mounted on a chestnut brown horse amidst the dense greenery of a forest. The horse stands at the edge of a tranquil river, its surface adorned with floating water lilies. Sunlight filters through the canopy, casting dappled shadows on the scene.

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about

Figure 10: More samples generated from LDGen.