LLM4GEN: Leveraging Semantic Representation of LLMs for Text-to-Image Generation

Mushui Liu[1,2*†], **Yuhang Ma**[2*], **Zhen Yang**[3], **Jun Dan**[1], **Yunlong Yu**[1‡], **Zeng Zhao**[2‡], **Zhipeng Hu**[2], **Bai Liu**[2], **Changjie Fan**[2]

[1]College of Information Science & Electronic Engineering, Zhejiang University
[2]Fuxi AI Lab, Netease Inc.
[3]The Hong Kong University of Science and Technology (Guangzhou)
lms@zju.edu.cn, mayuhang@corp.netease.com, zhen.yang@connect.hkust-gz.edu.cn, jundan@zju.edu.cn, yuyunlong@zju.edu.cn, {hzzhaozeng, zphu, hzliubai, fanchangjie}@corp.netease.com

Abstract

Diffusion models have exhibited substantial success in text-to-image generation. However, they often encounter challenges when dealing with complex and dense prompts involving multiple objects, attribute binding, and long descriptions. In this paper, we propose a novel framework called **LLM4GEN**, which enhances the semantic understanding of text-to-image diffusion models by leveraging the representation of Large Language Models (LLMs). It can be seamlessly incorporated into various diffusion models as a plug-and-play component. A specially designed Cross-Adapter Module (CAM) integrates the original text features of text-to-image models with LLM features, thereby enhancing text-to-image generation. Additionally, to facilitate and correct entity-attribute relationships in text prompts, we develop an entity-guided regularization loss to further improve generation performance. We also introduce DensePrompts, which contains 7,000 dense prompts to provide a comprehensive evaluation for the text-to-image generation task. Experiments indicate that LLM4GEN significantly improves the semantic alignment of SD1.5 and SDXL, demonstrating increases of 9.69% and 12.90% in color on T2I-CompBench, respectively. Moreover, it surpasses existing models in terms of sample quality, image-text alignment, and human evaluation.

Introduction

Recently, diffusion models (Song et al. 2020; Rombach et al. 2022; Huang et al. 2024; Shen et al. 2023; Wang et al. 2024; Yang et al. 2024b; He et al. 2024b; Tang et al. 2022b; Liu et al. 2023; Wang et al. 2025) have made significant progress in text-to-image (T2I) generation models, such as Imagen (Saharia et al. 2022), DALL-E (Betker et al. 2023), and Stable Diffusion (Rombach et al. 2022; Podell et al. 2023). However, they often encounter challenges in generating images given complex and dense prompt descriptions, such as attribute binding and multiple objects (Huang et al. 2023).

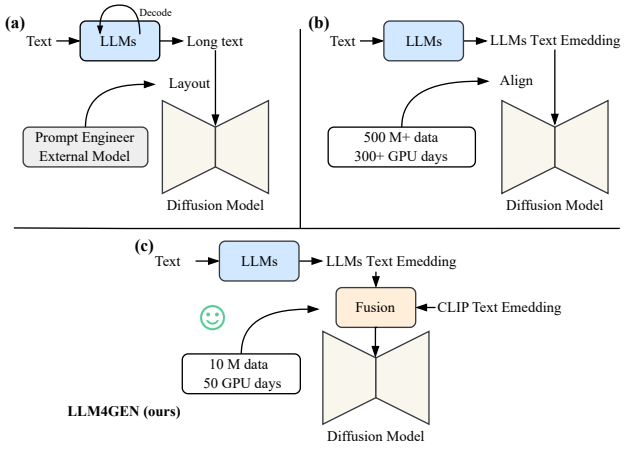


Figure 1: Architecture comparison between (a) LLM-guidance models (b) LLM-alignment models and (c) our proposed LLM4GEN.

With the emergence of powerful linguistic representations from Large Language Models (LLMs), there has been an increasing trend in leveraging LLMs to aid in T2I generation. Current methods mainly consist of two categories: LLM-guidance models (Yang et al. 2024a; Feng et al. 2022) and LLM-alignment models (Wu et al. 2023a; Hu et al. 2024; Esser et al. 2024; Zhao et al. 2024). LLM-guidance models harness the reasoning capability of LLMs and the Layout model to generate controllable images, as illustrated in Fig. 1 (a). However, these methods require separating LLMs from external models, resulting in redundancy in both inference time and the overall framework. While LLM-alignment models utilize LLMs to exploit the representational capacity, they demand substantial training data to align LLM representations with the diffusion model, as shown in Fig. 1 (b).

To address aforementioned challenges, we propose a novel framework named **LLM4GEN** that implicitly leverages the powerful semantic representations of **LLMs** to enhance the original text encoder **for** T2I **GEN**eration, as illustrated in Fig. 1 (c). Specifically, we design an efficient Cross-

*These authors contributed equally.
†This work was done during his internship at Fuxi Al Lab.
‡Corresponding Authors.

Figure 2: Image generation using concise and dense prompts, with colored text highlighting key entities or attributes, SDXL, Playgroundv2, Pixart-α, and our LLM4GEN$_{SDXL}$ (Zoom in for details).

Adapter Module (CAM) to implicitly integrate the semantic representation of LLMs with original text encoders that have limited representational capabilities, such as CLIP text encoder (Radford et al. 2021). We apply cross-attention to the representations of both encoder-only LLMs (*e.g.*, T5 (Raffel et al. 2020)) and decoder-only LLMs (*e.g.*, Llama (Zhang et al. 2024)) alongside CLIP text embeddings, and then concatenate fused embedding with original CLIP text embedding. CAM module significantly enhances the performance of T2I diffusion models while preserving the original text encoder representations, thereby reducing the need for extensive training costs. Additionally, we introduce an entity-guidance regularization loss that penalizes mismatches between the activation maps of entities and their corresponding attributes in the text, improving the model's ability to accurately comprehend and represent the main subjects in the generated images. As evidenced in Fig. 2, our proposed method exhibits strong performance in T2I generation.

To comprehensively assess the image generation capabilities of T2I models, we develop a comprehensive benchmark named **DensePrompts**, an extension of T2I-CompBench (Huang et al. 2023), which incorporates over 7,000 compositional prompts. The construction of this benchmark involves leveraging LLMs for complex text descriptions, followed by manual refinement. Results from performance metrics and human evaluations consistently demonstrate that LLM4GEN's representational capability surpasses other existing methods. Overall, our contributions are as follows:

- We propose a novel framework that leverages the powerful representational capabilities of LLMs to assist in text-to-image (T2I) generation. Specifically, we design a Cross-Adapter to integrate LLM representations and introduce an entity-guidance regularization loss to enhance semantic understanding.
- To assess performance with long-text prompts, we introduce DensePrompts, a benchmark designed to evaluate both aesthetic quality and image-text alignment.
- Our designed LLM4GEN can be seamlessly integrated into existing diffusion models like SD1.5 (Rombach et al. 2022) and SDXL (Podell et al. 2023). Experiments show that LLM4GEN exhibits superior performance in sample quality, image-text alignment, and human evaluation compared with existing state-of-the-art models.

Related Work

Large Language Models. Large language models (LLMs) (Chang et al. 2023) have shown powerful generalization ability in various NLP tasks. Recent LLMs, e.g., GPTs (Brown et al. 2020), LLaMA (Touvron et al. 2023), OPT (Zhang et al. 2022), PaLM (Chowdhery et al. 2022) are all equipped with billions of parameters, enabling the intriguing capability for in-context learning and demonstrating excellent zero-shot performance across various tasks. Certain Multi-modal LLMs (MLLMs) (Achiam et al. 2023; Zhu et al. 2023; Bai et al. 2023) have integrated visual and audio modalities, enhancing intelligent interactions with the help of LLMs. (Pang et al. 2024) shows that the frozen LLMs can further integrate visual understanding. Recent works (Lian et al. 2024; Sun et al. 2024; Yang et al. 2024a) use LLMs to create improved text prompts or bounding box layouts for high-quality text-to-image generation. However, these existing works only consider LLMs as simple condition generators, *e.g.*, text prompts or layout planning. In this paper, we harness the representation capabilities of LLMs to enhance text-to-image generation, emphasizing their significant representational power beyond simple text output.

Text-to-Image Diffusion Models. Text-to-image generation (Feng et al. 2024; He et al. 2024a; Shen et al. 2024b,a; Zhao et al. 2024; Shen et al. 2024c) aims to create images with given prompts. Diffusion models (Song and Ermon 2019, 2020; Song et al. 2020; Tang et al. 2022a) have demonstrated remarkable performance in image generation. These models use added Gaussian noise for a forward process and can generate diverse, high-quality images through an inverse process from random Gaussian noise. GLIDE (Nichol et al. 2022) utilizes CLIP (Radford et al. 2021) text encoder to enhance the image-text alignment. Latent Diffusion Models (LDMs) (Rombach et al. 2022) transfer the diffusion process from pixel to latent space. Recent models such as SD-XL (Podell et al. 2023), DALL-E 3 (Betker et al. 2023), and Dreambooth (Ruiz et al. 2023) have significantly enhanced image quality and text-image alignment using various perspectives, such as training strategies and scaling training data. Despite these notable advancements, generating high-fidelity images aligned with complex textual prompts remains challenging. In this paper, we propose LLM4GEN, which leverages the robust representation capabilities of LLMs to facilitate image generation.

Methodology

LLM4GEN

Framework The proposed LLM4GEN, which contains a Cross-Adapter Module (CAM) and the UNet, is illustrated in Fig. 3 (a). In this paper, we explore stable diffusion (Rombach et al. 2022; Podell et al. 2023) as the base text-to-image diffusion model, and the vanilla text encoder is from CLIP (Radford et al. 2021). LLM4GEN leverages the strong capability of LLMs to assist in text-to-image generation. The CAM extracts the representation of a given prompt via the combination of LLM and CLIP text encoder. The fused text

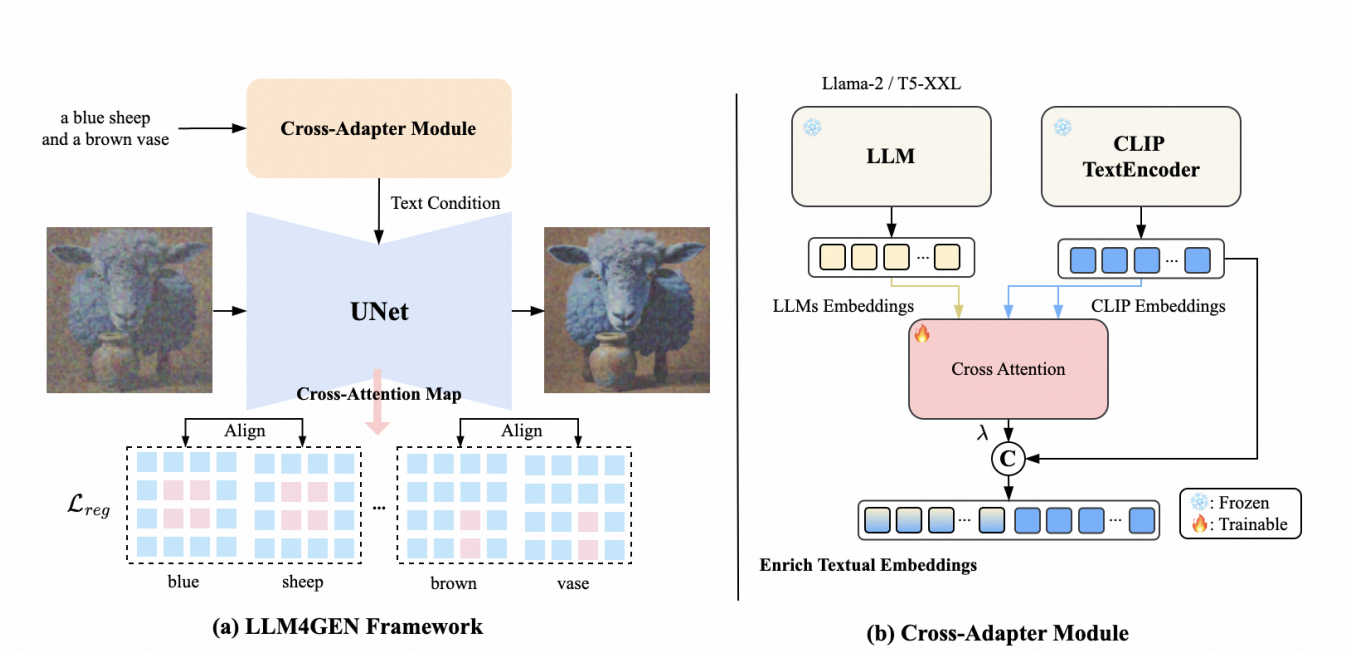


Figure 3: The overview of LLM4GEN. (a) Framework. (b) Cross-Adapter Module.

embedding is enhanced by leveraging the pre-trained knowledge of LLMs through the simple yet effective CAM. By feeding the fused text embedding, LLM4GEN iteratively denoises the latent vectors with the UNet and decodes the final vector into an image with the VAE.

Cross-Adapter Module The CAM connects the LLMs and the CLIP text encoder using a cross-attention layer, followed by concatenation with the representation of the CLIP text encoder. The last hidden state of the LLMs is extracted as LLMs feature c_l. The feature of CLIP text encoder is denoted as c_t, and we perform a cross-attention to fuse them:

$$Q = W_q(c_l), K = W_k(c_t), V = W_v(c_t) \tag{1}$$

$$c_l' = \text{CrossAttention}(Q, K, V) = \text{softmax}\left(Q \cdot K^T\right) \cdot V \tag{2}$$

where W_q, W_k, W_v are the trainable linear projection layers. The output embedding dimension is the same as that of CLIP text encoder. Then the final fused text embedding of the CAM is:

$$x = \text{CA}(x, \text{Concat}(\lambda \cdot c_l', c_t)) = \lambda \cdot \text{CA}(x, c_l) + \text{CA}(x, c_t) \tag{3}$$

where Concat denotes concatenation in the sequence dimension, and λ is the balance factor, x denotes the latent noise, CA is the cross-attention module within the UNet module. Overall, our designed Cross-Adapter Module implicitly facilitates the strong representation of LLMs with a residual fusion manner, without utilizing extensive training data and resources to condition the latent vectors on text embeddings. Notably, our LLM4GEN is compatible with both decoder-only and encoder-only LLMs and we evaluate on Llama-2 7B/13B (Touvron et al. 2023) and T5-XL (Brown et al. 2020) in further experiments.

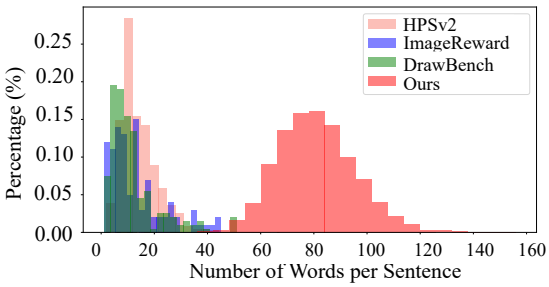


Figure 4: Statistic of DensePrompts benchmark compared with other benchmarks.

Entity-Guidance Regularization Loss Current text-to-image generation models (Rassin et al. 2024) often encounter confusion and omissions when generating multiple entities. We utilize a parser to analyze the prompt $\mathcal{P}$, extracting a set of attribute-entity pairs $\mathcal{S} = \sum_{i=1}^{N}\{a_i, e_i\}$, where e_i and a_i represent the entity name and its corresponding attribute, respectively, and N denotes the number of parsed pairs. Subsequently, we can calculate the active similarity map as:

$$\mathcal{A}_i = \text{softmax}\left(\frac{QK_i^T}{\sqrt{d}}\right) \tag{4}$$

where the query Q is derived from the latent representation, the key K_i is derived from the token embedding of p, and d is the latent dimension. $\mathcal{A}_a$ and $\mathcal{A}_o$ indicate the similarity maps for the attribute word and the entity, respectively. Subsequently, we impose a penalty on these similarity maps on all UNet layers as:

$$\mathcal{L}_{reg} = \frac{1}{N \cdot L}\sum_{i=1}^{N}\sum_{l=1}^{L}||\mathcal{A}_a^{i-l} - \mathcal{A}_o^{i-l}||^2 \tag{5}$$

where $||\cdot||$ represents the L2 distance, L is the layer numbers.

Overall, based on the framework described above, the training loss of LLM4GEN is formulated as:

$$\mathcal{L} = \mathbb{E}_{\varepsilon(x),\epsilon\sim\mathcal{N}(0,1),t}[\| \epsilon - \epsilon_\theta(z_t, t) \|_2^2] + \alpha \cdot \mathcal{L}_{reg} \quad (6)$$

where z_t can be obtained from the encoder $\mathcal{E}$, ans latent vectors from $p(z)$ can be decoded to images through the decoder $\mathcal{D}$. In this paper, we address the limited representation of CLIP as a text encoder by leveraging the capabilities of large language models (LLMs) to enhance the text encoder of the LDMs.

DensePrompts Benchmark

A comprehensive benchmark is crucial for evaluating the image-text alignment of generated images. Current benchmarks, *e.g.*, HPSv2 (Wu et al. 2023b), ImageReward (Xu et al. 2023), and DrawBench (Saharia et al. 2022), primarily consist of concise textual descriptions, are not comprehensive enough to describe a diverse range of objects. Thus, we introduce a new complicated benchmark called **DensePrompts**, comprising lengthy textual descriptions.

Initially, we collect 100 images from the Internet, comprising 50 real and 50 generated images, each with intricate details. Leveraging the robust image comprehension capabilities of GPT-4V (OpenAI 2023), we utilize it to provide detailed descriptions for these 100 images, encompassing object attributes and their relationships, thereby generating comprehensive prompts abundant in semantic details. We employ GPT-4 (Achiam et al. 2023) to produce massive long texts based on generated prompts mentioned above. DensePrompts provides more than 7,000 extensive prompts whose average word length is more than 40. Word statistics of DensePrompts are outlined in Fig. 4. To assess the performance, DensePrompts benchmark incorporates CLIP Score (Radford et al. 2021) and Aesthetic Score. Combining our proposed DensePrompts with T2I-CompBench, we establish a comprehensive evaluation in text-to-image generation.

Experiments

Experimental Details

Framework and Implementation Details In this paper, we explore LLM4GEN based on SD1.5 and SDXL, denoted as LLM4GEN$_{SD1.5}$ and LLM4GEN$_{SDXL}$. We utilize T5-XL and CLIP text encoder (CLIP ViT-L/14) as the text tower. The sequence length of the LLMs is set to 128. We use 10M text-image pairs collected from LAION-2B(Schuhmann et al. 2021) and Internet. Training is conducted on 8NVIDIA A100 GPUs with the learning rates of 2e-5 and 1e-5 for LLM4GEN$_{SD1.5}$ and LLM4GEN$_{SDXL}$, respectively. The batch size is set to 256 and 128. The training steps are set to 20k and 40k. Additionally, we further train LLM4GEN$_{SDXL}$ using 2M high-quality data with 1024 resolution. During inference, we utilize DDIM sampler (Song, Meng, and Ermon 2020) for sampling with 50 steps and the classifier free guidance scale to 7.5.

Evaluation Benchmarks We comprehensively evaluate our method via four primary benchmarks, *e.g.* **MSCOCO** (Lin et al. 2014), **T2I-CompBench** (Huang et al. 2023), our proposed **DensePrompts** benchmark, and **User Study**.

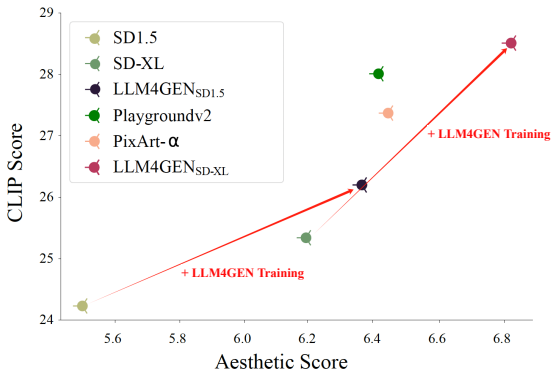


Figure 5: Aesthetic Score and CLIP Score (%) on DensePrompts benchmark.

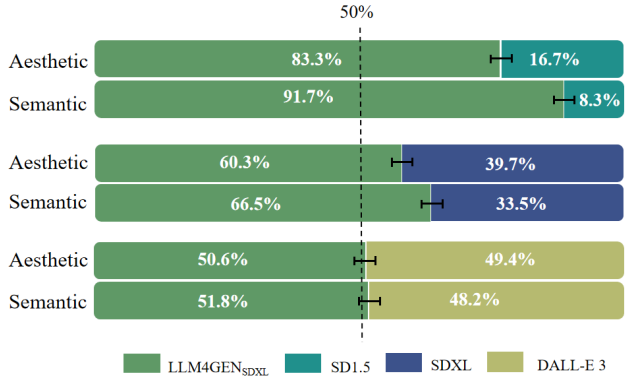


Figure 6: Results on user study regarding the sample quality and image-text alignment of different models.

Performance Comparisons and Analysis

Fidelity Assessment on MSCOCO Benchmark Experimental results on MSCOCO benchmark are shown in Tab. 2. LLM4GEN notably enhances the sample quality and image-text alignment, resulting in improvements of 1.79 and 0.54 on FID compared to SD1.5 and SDXL, respectively. Furthermore, we assess the performance of SD1.5 after extensive fine-tuning with the same training dataset. This modified version, SD1.5 (ft), surpasses the original SD1.5, yet LLM4GEN$_{SD1.5}$ still exhibits superior performance over SD1.5 (ft). This underscores the potent representation of our proposed LLM4GEN and its contribution to text-to-image generation.

Evaluation on T2I-CompBench. For T2I-CompBench comparison, we select the recent text-to-image generative models for comparison, e.g., Composable Diffusion, Structured Diffusion, Attn-Exct v2, GORS, DALLE 2, PixArt-α, ELLA$_{SDXL}$, SD1.5, and SDXL. Experimental results shown in Tab. 1 demonstrate the distinctive performance of LLM4GEN$_{SDXL}$ in T2I-CompBench evaluation, underlining its advancements in attribute binding, object relationship, and mastery in rendering complex compositions. LLM4GEN shows considerable improvement in color, shape, and texture, showcasing enhancements up to +12.90% in color, +5.16% in shape, and +14.49% in texture with SDXL, respectively. LLM4GEN$_{SDXL}$ also marks

Model	Attribute Binding			Object Relationship		Complex↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
Composable Diffusion (Liu et al. 2022)	40.63	32.99	36.45	8.00	29.80	28.98
Structured Diffusion (Feng et al. 2022)	49.90	42.18	49.00	13.86	31.11	33.55
Attn-Exct v2 (Chefer et al. 2023)	64.00	45.17	59.63	14.55	31.09	34.01
GORS (Huang et al. 2023)	66.03	47.85	62.87	18.15	31.93	33.28
DALL-E 2 (Ramesh et al. 2022)	57.50	54.64	63.74	12.83	30.43	36.96
PixArt-α (Chen et al. 2023)	68.86	55.82	70.44	20.82	31.79	41.17
SUR-Adapter* (Zhong et al. 2024)	38.93	36.73	39.21	13.21	30.81	31.43
ELLA$_{SDXL}$ (Hu et al. 2024)	72.60	56.34	66.86	22.14	30.69	-
SD1.5 (Rombach et al. 2022)	37.65	35.76	41.56	12.46	30.79	30.80
LLM4GEN$_{SD1.5}$ (Ours)	47.34	45.35	55.39	14.79	31.00	33.97
Δ (Margin)	+9.69	+9.59	+13.83	+2.33	+0.21	+3.17
SDXL (Podell et al. 2023)	63.69	54.08	56.37	20.32	31.10	40.91
LLM4GEN$_{SDXL}$ (Ours)	**76.59**	**59.24**	**70.86**	**24.12**	**32.10**	**42.19**
Δ (Margin)	+12.90	+5.16	+14.49	+3.80	+1.00	+1.28

Table 1: Evaluation results (%) on T2I-CompBench (Huang et al. 2023). The higher is better, and the best results are highlighted in bold. * denotes that we calculate the metrics using the official weights and code provided.

Method	**FID↓**	**IS↑**	**CLIP Score(%)↑**
SD1.5 (Rombach et al. 2022)	26.89	32.24	28.66
SD1.5 (ft)	25.48	33.53	29.10
LLM4GEN$_{SD1.5}$	**25.20**	**34.24**	**29.45**
SDXL (Podell et al. 2023)	24.75	34.91	30.10
LLM4GEN$_{SDXL}$	**24.21**	**35.10**	**30.91**

Table 2: Quantitative comparison on text-to-image generation models on the subset of MSCOCO (Lin et al. 2014).

considerable progress in both spatial and non-spatial evaluations, with 3.80% and 1.00% lift, respectively. Furthermore, when compared with PixArt-α, which employs T5-XL as its text encoder, LLM4GEN$_{SDXL}$ surpasses it in several aspects, such as a notable 7.73% lead in color metric. Moreover, LLM4GEN$_{SDXL}$ outperforms ELLA$_{SDXL}$. These results verify the potent synergy of LLMs representations in augmenting the sample quality and image-text alignment of diffusion models.

Evaluation on DensePrompts. We compare our LLM4GEN with PixArt-α, Playground v2, SD1.5, and SDXL on our DensePrompts benchmark. As shown in Fig. 5, LLM4GEN$_{SDXL}$ achieves the highest Aesthetic Score and CLIP Score among these models. PixArt-α outperforms SDXL due to its T5-XL text encoder for dense prompts. LLM4GEN excels in understanding and interpreting dense prompts, resulting in high-quality images with strong image-text alignment. This performance is attributed to the powerful representation of LLMs and the effective adaptation of the original CLIP text encoder via our CrossAdapter Module. We attribute this performance to the powerful representation of LLMs and the effective adaptation of the original CLIP text encoder via our CrossAdapter Module.

Quantitive Results. To thoroughly evaluate our proposed LLM4GEN framework, we present the qualitative results on the short prompts provided by PartiPrompts (Yu et al. 2022) in the first 4 columns and on the dense prompts provided by DensePrompts in the last 3 columns in Fig. 7. The results indicate that our proposed LLM4GEN$_{SD1.5}$ and LLM4GEN$_{SDXL}$ exhibit strong text-image alignment and superior dense prompt generation compared to the recent PixArt-α, especially in handling the multiple objects and attribute binding.

LLMs	Attribute Binding		
	Color ↑	Shape ↑	Texture ↑
SD1.5 (Rombach et al. 2022)	37.65	35.76	41.56
Llama-2/7B (Touvron et al. 2023)	43.21	40.12	48.91
Llama-2/13B (Touvron et al. 2023)	44.98	41.03	49.21
T5-XL (Raffel et al. 2020)	47.34	45.35	55.39

Table 3: Impact (%) of Different LLMs based on SD1.5.

User Study. We conduct the user study on various combinations of existing methods and LLM4GEN$_{SDXL}$. For each pairing, we assess two criteria: sample quality and image-text alignment. Users are tasked with evaluating the aesthetic appeal and semantic understanding of images with identical text to determine the superior one based on these assessment criteria. Subsequently, we compute the percentage scores for each model, as shown in Fig. 6. The results showcase our LLM4GEN$_{SDXL}$ exhibits comparative advantages over both SD1.5 and SDXL. Specifically, LLM4GEN$_{SDXL}$ achieves 60.3% and 66.5% higher voting preferences compared to SDXL in terms of Aesthetic and Semantic, respectively. Notably, LLM4GEN$_{SDXL}$ also delivers competitive results when compared to DALL-E 3.

Ablation Studies

Impact of Different LLMs. The analysis encompasses a comparative evaluation between base SD1.5 and the enhancements achieved through the integration of Llama-

Figure 7: A comparative analysis of LLM4GEN and other state-of-the-art diffusion models using PartiPrompts (Yu et al. 2022) and our proposed DensePrompts as prompts.

Module	Attribute Binding		
	Color ↑	Shape ↑	Texture ↑
(1) SD1.5	37.65	35.76	41.56
(2) SD1.5 finetune-CLIP	38.76	36.85	42.56
(3) MLP or CrossAttention	39.25	37.68	42.89
(4) MLP + Concat	40.24	38.23	44.39
(5) CrossAttention + Concat	42.31	39.43	46.21
(6) CLIP as Q and LLM as KV	43.69	42.82	49.56
(7) w/o Reg.	46.10	44.31	54.81
(8) Ours	**47.34**	**45.35**	**55.39**

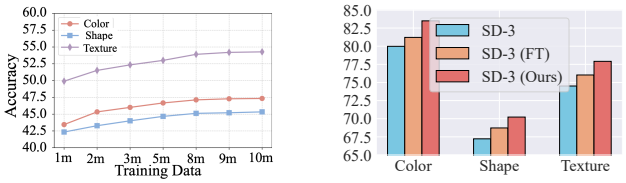
Table 4: Impact (%) of the designed Cross-Adapter Module.

2/7B, Llama-2/13B, and T5-XL. As depicted in Tab. 3, the inclusion of any LLM improves upon the performance of SD1.5. Importantly, Llama-v2/13B outperforms Llama-v2/7B, demonstrating that LLMs with greater capacity excel in extracting more nuanced semantic embeddings. Furthermore, when compared to decoder-only LLMs, T5-XL encoder demonstrates advantages in semantic comprehension, confirming its superior suitability for enhancing text-to-image generation.

Impact of Modules. Due to limited computing sources, we evaluate the impact of various architectural enhancements on SD1.5, as outlined in Tab. 4. Our configurations explore different methods for integrating LLMs embeddings: (1) the baseline SD1.5 model, (2) SD1.5 finetune-CLIP, the result of fine-tuning the original text-encoder of SD1.5, (3) MLP or CrossAttention, which utilizes a simple linear layer or cross-attention layer to transform LLM embeddings, (4) MLP + Concat, representing a process where LLMs embeddings are projected to the same dimension as the original text embeddings before concatenation, (5) CrossAttention + Concat, (6) CLIP as Q and LLM as KV, refering to converting the position of Q and KV in CAM, (7) w/o Reg., referring to that training withour regularization loss. Results show that configuration (2) indeed brings an improvement of the original SD1.5, yet, due to the limited semantic representation of CLIP, the results still remain subpar. Interestingly, simply concatenating the original text embeddings (configuration 3 & 4) provides a significant boost over base SD1.5. This suggests that direct representation alignment between LLMs and the latent vector is challenging, and enhancing the original text embeddings with LLM embeddings is sufficient to improve image-text alignment. In our LLM4GEN,

Method	#Images (M ↓)	#GPU days (↓)	Performance (↑)
PixArt-α (Chen et al. 2023)	25	753	68.86
ParaDiffusion (Wu et al. 2023a)	500	392	-
ELLA$_{SDXL}$ (Hu et al. 2024)	30	112	72.60
LLM4GEN$_{SDXL}$	10	50	**76.59**

Table 5: Training resources comparison, including the scale of training data and computing cost.



(a) Training Data. (b) Compatibility with SD-3.

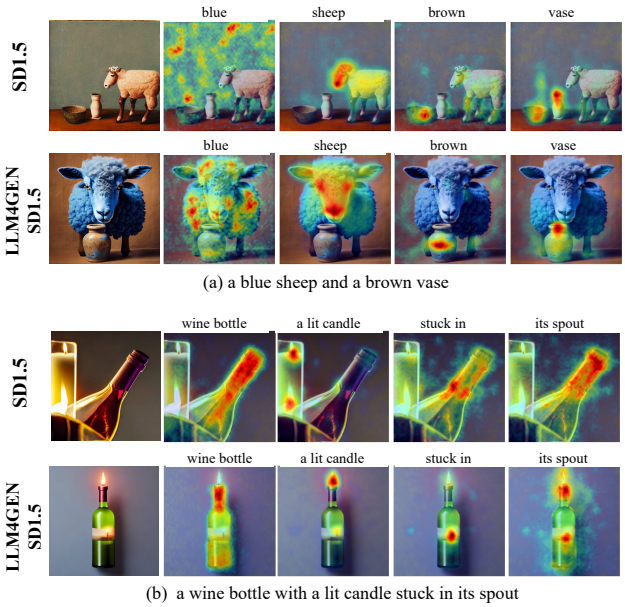
Figure 8: (a) Training data analysis and (b) Compatibility with SD-3.

the LLM representation is employed as Q, while the original text encoder serves as K and V. We also examine the impact of rearranging the position of Q and KV in the CAM module. The results, as demonstrated in (6), indicate that our LLM4GEN (8) exceeds it, showcasing a 3.65% enhancement in color. This emphasizes the substantial benefits of incorporating our Cross-Adapter Module to enrich the representation of the original text encoder and the image-text alignment of generated images.

Further Analysis

Scaling Analysis. As illustrated in Fig. 8a, we conduct an extensive analysis of the scalability of our proposed LLM4GEN model. The results conclusively demonstrate that as the quantity of training data increases, the performance of our model exhibits consistent and significant growth, thereby confirming its scalability. However, increasing the dataset scale from 5M to 10M resulted in minimal performance improvement on the generated images. Consequently, we use 10M text-image pairs for training our LLM4GEN. Notably, we demonstrate that incorporating the Llama-3 whose model size is 8B into SD-3-medium (SD-3) can bring further performance improvements, showcasing the flexibility of our method, as shown in Fig. 8b. With the same dataset, our approach consistently outperforms both direct fine-tuning and the original SD-3 model. This compatibility offers great meaningful contributions to the community, enabling the integration of different LLMs to enhance existing text-to-image generation models.

Training Efficiency. When evaluating the effectiveness of integrating LLMs into text-to-image generation models, LLM4GEN$_{SDXL}$ stands out for its remarkable efficiency and performance. LLM4GEN achieves significant reductions in both training data requirements and computational costs. It utilizes only 10 million data, a 66% reduction compared to ELLA, and demands merely 50 GPU days for training, drastically lower than PixArt-α (25 million data, 753 GPU days) and ParaDiffusion (500 million data, 392 GPU days). Despite this, LLM4GEN$_{SDXL}$ achieves a superior color metric performance of 73.29%. This notable difference underscores LLM4GEN's ability to substantially reduce both training data and computational costs while establishing a new standard for performance efficiency.



(a) a blue sheep and a brown vase

(b) a wine bottle with a lit candle stuck in its spout

Figure 9: Cross-attention visualization (Tang et al. 2022c) for two generated images. The two rows are SD1.5 and LLM4GEN$_{SD1.5}$, respectively.

Cross-attention Visualization. Fig. 9 shows the cross-attention visualization of SD1.5 and LLM4GEN$_{SD1.5}$, respectively. The heatmaps reveal that our proposed LLM4GEN method demonstrates a superior ability to capture relationships between attributes, such as "blue" and "sheep," as illustrated in Fig. 9 (a). We attribute this enhanced capability to the increased semantic richness afforded by the robust representations of LLMs.

Conclusion

In this paper, we propose LLM4GEN, an end-to-end text-to-image generation framework. Specifically, we design an efficient Cross-Adapter Module to leverage the powerful representation of LLMs, thereby enhancing the original text representation of diffusion models. Despite using fewer training data and computational resources, LLM4GEN outperforms current state-of-the-art text-to-image diffusion models in sample quality and image-text alignment. To optimize consistency in entity-attribute relationships of generated images, we design an entity-guided regularization loss. Additionally, we introduce the DensePrompts benchmark to promote the generation of images with dense information and provide a comprehensive evaluation framework. Extensive experiments have shown that our proposed method achieves competitive performance.

Acknowledgments

This research was supported in part by the Key R&D Program of Zhejiang Province, China 2023C01043, 2024C01G1752215, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, Science and Technology Innovation 2025 Major Project of Ningbo (2023Z236), the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM TOG*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *ICLR*.

Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2024. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2024a. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*.

He, X.; Xu, G.; Zhang, B.; Chen, H.; Cui, Y.; and Guo, D. 2024b. DiffCalib: Reformulating Monocular Camera Calibration as Diffusion-Based Dense Incident Map Generation. *arXiv preprint arXiv: 2405.15619*.

Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; and Yu, G. 2024. ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment. *arXiv preprint arXiv:2403.05135*.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In *ICCV*.

Huang, Q.; Fu, S.; Liu, J.; Jiang, H.; Yu, Y.; and Song, J. 2024. Resolving multi-condition confusion for finetuning-free personalized image generation. *arXiv preprint arXiv:2409.17920*.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2024. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *Transactions on Machine Learning Research*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *ECCV*, 423–439.

Liu, Y.; Zhang, J.; Peng, D.; Huang, M.; Wang, X.; Tang, J.; Huang, C.; Lin, D.; Shen, C.; Bai, X.; et al. 2023. Spts v2: single-point scene text spotting. *IEEE T-PAMI*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*.

OpenAI. 2023. GPT-4V(ision) System Card.

Pang, Z.; Xie, Z.; Man, Y.; and Wang, Y.-X. 2024. Frozen transformers in language models are effective visual encoder layers. In *ICLR*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 8748–8763.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1): 5485–5551.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *NeurIPS*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR*, abs/2111.02114.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; Shu, X.; Du, X.; and Tang, J. 2023. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In *ACM MM*.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2024b. Boosting consistency in story visualization with rich-contextual conditional diffusion models. *arXiv preprint arXiv:2407.02482*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024c. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *ICLR*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *CoRR*, abs/2010.02502.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; and Wang, X. 2024. Emu: Generative pretraining in multimodality. In *ICLR*.

Tang, J.; Qian, W.; Song, L.; Dong, X.; Li, L.; and Bai, X. 2022a. Optimal boxes: boosting end-to-end scene text recognition by adjusting annotated bounding boxes via reinforcement learning. In *ECCV*, 233–248.

Tang, J.; Zhang, W.; Liu, H.; Yang, M.; Jiang, B.; Hu, G.; and Bai, X. 2022b. Few could be better than all: Feature sampling and grouping for scene text detection. In *CVPR*, 4563–4572.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Ture, F. 2022c. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, A.-L.; Shan, B.; Shi, W.; Lin, K.-Y.; Fei, X.; Tang, G.; Liao, L.; Tang, J.; Huang, C.; and Zheng, W.-S. 2025. ParGo: Bridging Vision-Language with Partial and Global Views. In *AAAI*.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2024. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. *arXiv preprint arXiv:2406.07209*.

Wu, W.; Li, Z.; He, Y.; Shou, M. Z.; Shen, C.; Cheng, L.; Li, Y.; Gao, T.; Zhang, D.; and Wang, Z. 2023a. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023b. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*.

Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; and Cui, B. 2024a. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*.

Yang, Z.; Ding, G.; Wang, W.; Chen, H.; Zhuang, B.; and Shen, C. 2024b. Object-aware inversion and reassembly for image editing. In *ICLR*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; and Qiao, Y. 2024. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhao, S.; Hao, S.; Zi, B.; Xu, H.; and Wong, K.-Y. K. 2024. Bridging Different Language Models and Generative Vision Models for Text-to-Image Generation. In *ECCV*.

Zhong, S.; Huang, Z.; Wen, W.; Qin, J.; and Lin, L. 2024. SUR-adapter: Enhancing Text-to-Image Pre-trained Diffusion Models with Large Language Models. In *ACM MM*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.