ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment

Xiwei Hu*, Rui Wang*, Yixiao Fang*, Bin Fu*, Pei Cheng, and Gang Yu**

Tencent
xiweihu@outlook.com, {raywwang, yixiaofang, brianfu, peicheng}@tencent.com, skicy@outlook.com
https://ella-diffusion.github.io

Abstract. Diffusion models have demonstrated remarkable performance in the domain of text-to-image generation. However, most widely used models still employ CLIP as their text encoder, which constrains their ability to comprehend dense prompts, encompassing multiple objects, detailed attributes, complex relationships, long-text alignment, *etc.* In this paper, we introduce an **E**fficient **L**arge **L**anguage Model **A**dapter, termed **ELLA**, which equips text-to-image diffusion models with powerful Large Language Models (LLM) to enhance text alignment *without training of either U-Net or LLM*. To seamlessly bridge two pre-trained models, we investigate a range of semantic alignment connector designs and propose a novel module, the Timestep-Aware Semantic Connector (TSC), which dynamically extracts timestep-dependent conditions from LLM. Our approach adapts semantic features at different stages of the denoising process, assisting diffusion models in interpreting lengthy and intricate prompts over sampling timesteps. Additionally, ELLA can be readily incorporated with community models and tools to improve their prompt-following capabilities. To assess text-to-image models in dense prompt following, we introduce Dense Prompt Graph Benchmark (DPG-Bench), a challenging benchmark consisting of 1K dense prompts. Extensive experiments demonstrate the superiority of ELLA in dense prompt following compared to state-of-the-art methods, particularly in multiple object compositions involving diverse attributes and relationships.

Keywords: Diffusion Models · Large Language Models · Text-Image Alignment

1 Introduction

In recent years, significant advancements have been made in text-to-image generation based on diffusion models. These models [7, 40, 43, 45, 47] are capable of generating text-relevant images with high aesthetic quality, thereby driving the development of open-source community models and downstream tools. To

* Equal Contribution
** Corresponding Author

Fig. 1: Comparison to SDXL [40] and DALL-E 3 [7]. The prompts originate from PartiPrompts [60] (colored text denotes critical entities or attributes).

adhere to prompt instructions, various existing models, especially [40, 43, 45], employ the pre-trained CLIP [41] model as a text encoder, which is trained on images with predominantly short text pairs. However, these models encounter difficulties in handling long dense prompts, particularly when the text describes multiple objects along with their distinct attributes and relationships. Some models [47,55] investigate the incorporation of powerful Large Language Models (LLM), such as T5 [42] and LLaMA-2 [52], with diffusion models to achieve a deeper level of language understanding in text-to-image generation. Imagen [47] first demonstrates that text features from LLMs pre-trained on text-only corpora are remarkably effective in enhancing text alignment for text-to-image synthesis. Nonetheless, current models [12, 47, 55] that employ LLM as a text encoder

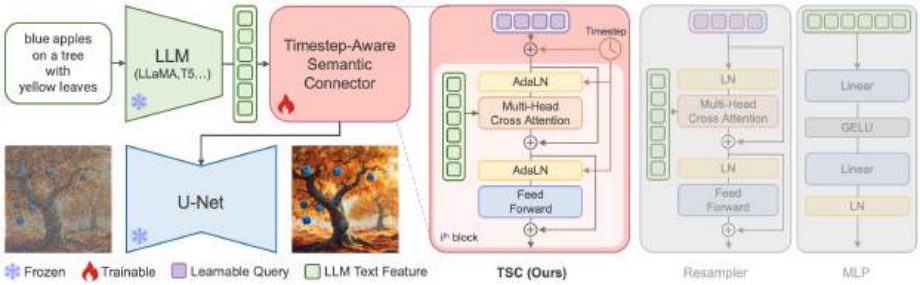

Fig. 2: The overview of ELLA. *Left:* The pipeline of our architecture. *Right:* The details of our TSC design alongside potential alternatives for connectors. We have conducted ablation studies on different connectors and finally selected TSC.

necessitate the full training of U-Net [46], and ParaDiffusion [55] even fine-tunes the pre-trained LLM. Aside from consuming vast computational resources, these models are difficult to integrate with the burgeoning community models and downstream tools [59, 61].

To address the dense prompt understanding limitations of CLIP-based diffusion models, we propose a novel approach named ELLA, which incorporates powerful LLM in a lightweight and efficient manner. The architecture of ELLA is illustrated in Fig. 2. Given pre-trained LLM and U-Net, we explore various semantic alignment connectors and train an adaptive connector, proposed as the Timestep-Aware Semantic Connector (TSC), on text-image pair data rich in information density. As observed by [5, 16, 24], diffusion models typically predict low-frequency content during the initial stages of the denoising process and subsequently concentrate on high-frequency details towards the final stage. Consequently, we anticipate our alignment connector to initially extract text features at the low-frequency semantic level, which corresponds to the main objects and layout described in the prompt. Conversely, during the latter stages of denoising, we expect TSC to extract high-frequency semantics, which pertain to detailed attributes. The architectural design of our TSC is based on the resampler [4], and it instills temporal dependency by integrating the timestep in the Adaptive Layer Normalization [38, 39]. To facilitate TSC in dense information comprehension, we assemble highly informative text-image pair datasets, the captions of which are generated by state-of-the-art Multi-modal Language Language Models (MLLM). Once trained, TSC can seamlessly integrate community models and downstream tools such as LoRA [26] and ControlNet [61], improving their text-image alignment.

Additionally, we introduce the Dense Prompt Graph Benchmark (DPG-Bench), a comprehensive dataset consisting of 1,065 lengthy, dense prompts, designed to assess the intricate semantic alignment capabilities of text-to-image models. In contrast to earlier benchmarks [14, 27, 60], DPG-Bench encompasses dense prompts that describe multiple objects, each characterized by a variety of attributes and relationships. This benchmark also facilitates automatic evaluation using state-of-the-art MLLM. Extensive experiments on T2I-CompBench [27]

and DPG-Bench demonstrate the superior semantic alignment of ELLA compared to existing SOTA T2I models [7, 40, 43, 45, 47].

Our key contributions include:

- We propose a novel lightweight approach ELLA to equip existing CLIP-based diffusion models with powerful LLM. Without training of U-Net and LLM, ELLA improves prompt-following abilities and enables long dense text comprehension of text-to-image models.
- We design a Timestep-Aware Semantic Connector (TSC) to extract timestep-dependent conditions from the pre-trained LLM at various denoising stages. Our proposed TSC dynamically adapts semantics features over sampling time steps, which effectively conditions the frozen U-Net at distinct semantic levels.
- We introduce the DPG-Bench, comprising 1,065 lengthy, dense prompts characterized by a multitude of attributes and relationships. Experimental results from user studies corroborate that the proposed evaluation metrics are highly correlated with human perception.
- Extensive results show that our ELLA exhibits superior performance in text alignment compared to existing state-of-the-art models, and significantly enhances the prompt-following capabilities of community models and downstream tools.

2 Related Work

2.1 Text-to-Image Diffusion Models.

Diffusion-based text-to-image models have exhibited remarkable enhancements in generating high-fidelity and diverse images. These models need powerful text encoders to comprehend detailed image descriptions. GLIDE [37], LDM [45], DALL-E 2 [43] and Stable Diffusion [40, 45] employ the pre-trained CLIP [41] model to extract text embeddings. Imagen [47], Pixart-α [12] and DALL-E 3 [7] use pre-trained large language models (e.g., T5 [42]) as text encoders, demonstrating that language text features exhibit a better understanding of text. eDiff-I [5] and EMU [17] use both CLIP and T5 embeddings as conditions. ParaDiffusion [55] proposes fine-tuning the LLaMA-2 [52] model during diffusion model training and using the fine-tuned LLM text features as the condition. We equip the pre-trained CLIP-based models with LLM [42, 52, 62] to enhance prompt following ability with the help of TSC.

2.2 Compositional Text-to-Image Diffusion models.

Various methods of compositional text-to-image diffusion models have been explored to better adhere to complex prompts. Some works [6, 11, 13, 20, 28, 31, 35, 44, 56] attempt to manipulate cross-attention maps or latents according to spatial or semantic constraints in the prompt. However, these methods are contingent upon the interpretability of the base models and can only achieve coarse

and suboptimal control [10, 25, 32, 58]. Another potential solution [18, 27, 50, 57] is to leverage image understanding feedback as reward to fine-tune the text-to-image models. These approaches are potentially constrained by the limitations in CLIP text understanding ability. Some studies [23, 63] employ LLMs to enhance the prompt or its corresponding embedding. [15, 21, 22, 32, 54, 58] harness the planning and reasoning ability of LLMs [3, 52] to deconstruct prompts into multiple regional descriptions, serving as condition to guide the image generation process. Our method enhances the base model's ability to follow prompts and can be seamlessly integrated with training-free methods [13, 20, 58].

3 Method

3.1 Architecture Design

To leverage the language understanding capability of LLM and the image generation potential of diffusion models and bridge them effectively, we design our ELLA as depicted in Fig. 2. We consider the pre-trained LLM, such as T5 [42], TinyLlama [62] and LLaMA-2 [52], as the text encoder, which provides the comprehensive text feature to condition image generation. A carefully designed Timestep-Aware Semantic Connector (TSC) receives the text feature with arbitrary length as well as the timestep embedding, and outputs fixed-length semantic queries. These semantic queries are used to condition noisy latent prediction of the pre-trained U-Net through cross-attention. To improve the compatibility and minimize the training parameters, we leave both the text encoder of Large Language Models as well as the U-Net and VAE components frozen. The only trainable component is consequently our lightweight TSC module.

Text Encoder. ELLA is compatible with any state-of-the-art Large Language Models as text encoder, and we have conducted experiments with various LLMs, including T5-XL [42], TinyLlama [62], and LLaMA-2 13B [52]. The last hidden state of the language models is extracted as the comprehensive text feature. The text encoder is frozen during the training of ELLA. The detailed performance comparison of different Large Language Models is given in Sec. 5.3.

Timestep-Aware Semantic Connector (TSC). This module interacts with the text features to facilitate improved semantic conditioning during the diffusion process. We investigate various network designs that influence the capability to effectively transfer semantic understanding.

- *MLP.* Following widely used converter design in LLaVA [34], we apply a similar MLP to map the text feature to image generation condition space.
- *Resampler.* As we connect two different modalities, MLP may not be sufficient enough to fuse and transfer the text feature. In addition, MLP is not a flexible design facing variant-length input and is difficult to scale up. We follow the Perceiver Resampler design from [4] to learn a predefined number of latent input queries with transformer-based blocks. The latent input queries interact with frozen text features through cross-attention layers, which allows the module to tackle input text of arbitrary token length.

- *Resampler with timestep using Adaptive Layer Norm (AdaLN).* While keeping the text encoder and the diffusion model frozen, we expect our connector to provide highly informative conditions for noise prediction. Inspecting a given prompt for image generation, we notice that certain words describe the primary objects and corresponding attributes, while others may delineate details and image style. It is also observed that during image generation, the diffusion model initially predicts the main scene and subsequently refines the details [24]. This observation inspires us to incorporate timestep into our resampler, allowing for the extraction of dynamic text features to better condition the noise prediction along with various diffusion stages. Our main results are based on this network design.
- *Resampler with timestep using AdaLN-Zero.* AdaLN-Zero is another possible design to introduce timestep, which has been demonstrated more effective in DiT [38]. However, this design is not the best choice in our framework and the experimental details can be found in Sec. 5.3.

3.2 Dataset Construction

Most diffusion models adopt web datasets [9,48] as the training dataset. However, the captions in these datasets, which are all alt-texts, often include overly brief or irrelevant sentences. This leads to a low degree of image-text correlation and a scarcity of dense semantic information. To generate long, highly descriptive captions, we apply the state-of-the-art MLLM CogVLM [53] as auto-captioner to synthesize image descriptions with the main objects and surroundings while specifying the corresponding color, texture, spatial relationships, *etc.* We show the vocabulary analysis [8] of a subset of LAION [48] and COYO [9] in Tab. 1, as well as our constructed CogVLM version. It can be observed that the alt-text in LAION/COYO contains significantly less information compared to the caption annotated by CogVLM. The latter features a greater number of nouns, adjectives, and prepositions, making it more descriptive in terms of multiple objects, diverse attributes, and complex relationships. We filter images collected from LAION and COYO with an aesthetic score [2] over 6 and a minimum short edge resolution of 512 pixels. We annotated in total 30M image captions using CogVLM, which improves the matching between images and text and increases the information density of the captions. To augment the diversity of prompt formats and improve the quality of generated images, we further incorporate 4M data from JourneyDB [51] with their original captions.

4 Benchmark

Current benchmarks have not considered evaluating the generation models' ability to follow dense prompts. The average token length of prompts tokenized by CLIP tokenizer in previous benchmarks, such as T2I-CompBench [27] and PartiPrompts [60], is about 10-20 tokens. Moreover, these benchmarks are not comprehensive enough to describe a diverse range of objects, as they contain limited words in each prompt and lack sufficient diversity of nouns throughout the

Table 1: Dataset information. All the numbers in the table are the **average** results of each text in the dataset. Token statistics are calculated by CLIP tokenizer. The abbreviation is derived from NLTK. **NN**: noun, including singulars, plurals and proper nouns. **JJ**: adjective, including comparatives and superlatives. **RB**: adverb, including comparatives and superlatives. **IN**: preposition and subordinating conjunctions.

Dataset	Words	NN	JJ/RB	IN	Tokens
LAION	9.81	3.59	0.70	1.87	11.88
LAION-CogVLM	49.87	15.51	8.06	6.26	62.33
COYO	9.83	3.60	0.65	1.91	11.89
COYO-CogVLM	50.71	15.71	8.06	6.38	63.05

Table 2: Benchmark information. **#DN** means the total distinct nouns in the benchmarks. All the numbers in the table are the **average** results except for #DN. Token statistics are calculated by CLIP tokenizer.

Benchmarks	#DN	Words	NN	JJ/RB	IN	Tokens
T2I-CompBench	1447	9.60	3.40	1.36	0.87	12.65
PartiPrompts	1421	9.11	3.45	0.94	1.36	12.20
DSG-1k	2004	17.13	6.35	2.09	2.41	22.56
DPG-Bench (ours)	4286	67.12	20.90	11.59	9.07	83.91

entire benchmark, as listed in Tab. 2. Therefore, we present a more comprehensive benchmark for dense prompts, called **Dense Prompt Graph Benchmark (DPG-Bench)**. Compared to previous benchmarks, as shown in Fig. 3, DPG-Bench provides longer prompts containing more information.

We gather source data from COCO [33], PartiPrompts [60], DSG-1k [14] and Object365 [49]. For the data originating from the first three sources, we create long dense prompts based on the original short prompts. In the case of Object365, we randomly sampled 1-4 objects for each prompt according to its main category and subcategory, subsequently generating prompts with sampled objects. Given the source data, we instruct GPT-4 to specify the details of the scene, the attributes as well as the corresponding relationship of objects, creating long dense prompts with rich semantic information. All prompts are generated automatically by GPT-4 and verified by humans. With constructed dense prompts, we follow the pipeline of DSG, leveraging GPT-4 once more to generate corresponding tuple categories, questions, and graphs. As shown

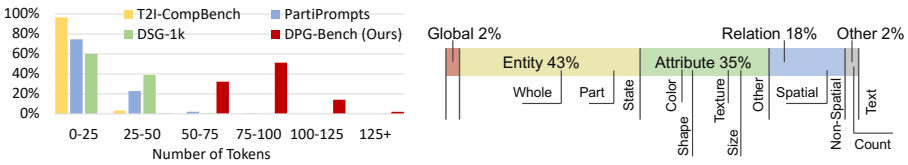

Fig. 3: DPG-Bench Information. *Left:* The token distribution of DPG-Bench and other benchmarks. *Right:* the level-1 categories and the level-2 categories.

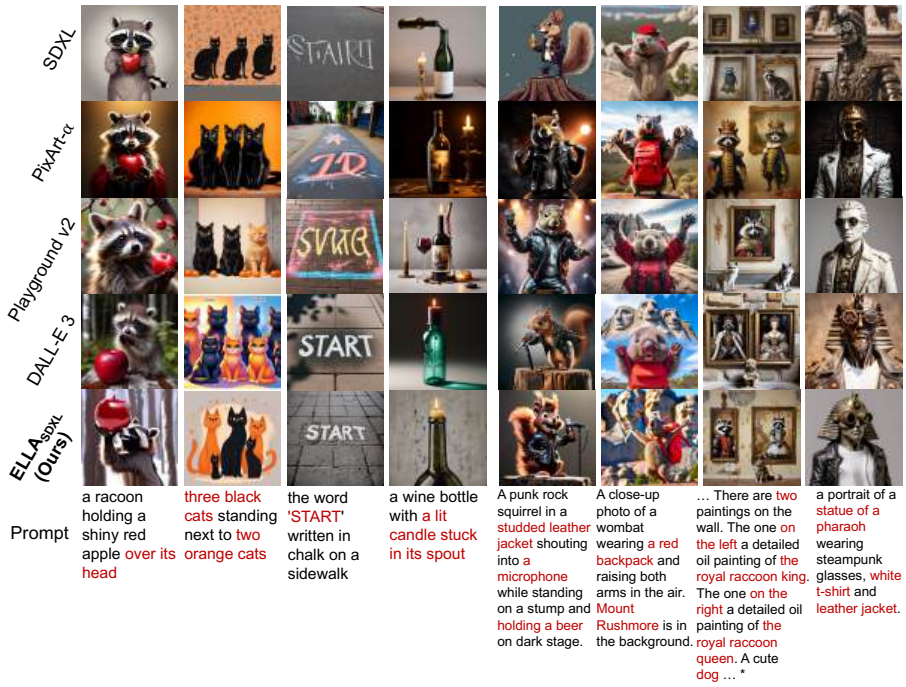

Fig. 4: The comparison between ELLA, SDXL, PixArt-α, Playground v2 [30] and DALL-E 3. The left four columns only contain 1 or 2 entities, but the right four correspond to dense prompts with more than 2 entities. All prompts originate from PartiPrompts. The complete prompt with the mark * is written in the footnote.[1]

in Fig. 3, our benchmark provides a two-level category scheme, with 5 level-1 categories and 13 level-2 categories.

In conducting the assessment, each model is expected to generate 4 images for each given prompt. Subsequently, mPLUG-large [29] is employed as the adjudicator to evaluate the generated images according to the designated questions. As for the score computation, our benchmark adheres to the principles of DSG. Ultimately, the mean score of a series of questions pertaining to a single prompt constitutes the prompt score, while the mean score of all prompt scores represents the DPG-Bench score.

5 Experiments

5.1 Implementation Details

Training Details. We employ the encoder of T5-XL [42], a 1.2B model for text feature extraction. To be more compatible with models in the open-source community, we use SDv1.5 [45] and SDXL [40] as base models for training. We set

[1] a wall in a royal castle. There are two paintings on the wall. The one on the left a detailed oil painting of the royal raccoon king. The one on the right a detailed oil painting of the royal raccoon queen. A cute dog looking at the two paintings, holding a sign saying 'plz conserve'

the training length of extracted text tokens as 128 to handle complex scene understanding in dense captions. Our models are trained on 34M image-text pairs with 512 resolution for text alignment. We further train our ELLA$_{SDXL}$ on 100K high-quality data with 1024 resolution for aesthetic improvements. The AdamW optimizer [36] is used with a weight decay of 0.01, a constant 1e-4 learning rate for ELLA$_{SDv1.5}$ and 1e-5 for ELLA$_{SDXL}$. The final model is trained on 8 40G A100 for approximately 7 days for the ELLA$_{SDv1.5}$ and 14 days for ELLA$_{SDXL}$. Our ELLA$_{SDXL}$ costs less than 80% training time compared to PixArt-α [12] (753 A100 GPU days).

Table 3: Evaluation on T2I-CompBench with short compositional prompts. The **bold** score denotes the best performance. Higher score for better performance.

Model	Attribute Binding			Object Relationship	
	Color	Shape	Texture	Spatial	Non-Spatial
SD v1.4	0.3765	0.3576	0.4156	0.1246	0.3079
SD v2	0.5065	0.4221	0.4922	0.1342	0.3096
Composable v2	0.4063	0.3299	0.3645	0.0800	0.2980
Structured v2	0.4990	0.4218	0.4900	0.1386	0.3111
Attn-Exct v2	0.6400	0.4517	0.5963	0.1455	0.3109
GORS	0.6603	0.4785	0.6287	0.1815	**0.3193**
DALL-E2	0.5750	0.5464	0.6374	0.1283	0.3043
PixArt-α	0.6886	0.5582	**0.7044**	0.2082	0.3179
SD v1.5	0.3750	0.3724	0.4159	0.1204	0.3088
ELLA$_{SDv1.5}$	0.6911	0.4938	0.6308	0.1867	0.3062
SDXL	0.6369	0.5408	0.5637	0.2032	0.3110
ELLA$_{SDXL}$	**0.7260**	**0.5634**	0.6686	**0.2214**	0.3069

5.2 Performance Comparison and Analysis

Alignment Assessment. To evaluate our model on short compositional prompts, we first conduct experiments on a subset of T2I-CompBench [27] to assess the alignment between the generated images and text conditions in attribute binding and object relationships. We compare our models with recent Stable Diffusion [45] v1.4, v1.5, v2 and XL [40] model, Composable v2 [35], Structured v2 [19], Attn-Exct v2 [11], GORS [27], PixArt-α [12] and DALL-E 2 [45]. As shown in Tab. 3, our ELLA$_{SDv1.5}$ performance significantly surpasses its based model SD v1.5 [45] and is even comparable to SDXL in some subset of evaluation categories. The ELLA$_{SDXL}$ performs better than SDXL and the fully-tuned PixArt-α in most categories. It is observed that the MLLM caption model is highly sensitive to information such as color and texture in images. Therefore, such data greatly contributes to the learning of these attributes.

Dense Prompt Following. To further measure a model's proficiency in following longer prompts laden with dense information, we test different models on our proposed DPG-Bench to evaluate semantic alignment for complex texts. As

Table 4: Evaluation results on DPG-Bench. Average score is the graph score based on the rule of DSG and the larger score is better. The other scores are the average of all questions in one category. The VQA answers are generated by mPLUG-large [29]. **#Params** denotes trainable parameters. Higher score for better performance. The **bold** score for the best and the **bold underlined** score for the second best.

Model	#Params	Average	Global	Entity	Attribute	Relation	Other
SD v2	0.86B	68.09	77.67	78.13	74.91	80.72	80.66
PixArt-α	0.61B	71.11	74.97	79.32	78.60	82.57	76.96
Playground v2	2.61B	74.54	83.61	79.91	82.67	80.62	81.22
DALL-E 3	-	**83.50**	90.97	89.61	88.39	90.58	89.83
SD v1.5	0.86B	63.18	74.63	74.23	75.39	73.49	67.81
ELLA$_{SDv1.5}$	0.07B	74.91	84.03	84.61	83.48	84.03	80.79
SDXL	2.61B	74.65	83.27	82.43	80.91	86.76	80.41
ELLA$_{SDXL}$	0.47B	**80.23**	85.90	85.34	86.67	86.16	87.41

depicted in Tab. 4, our model has the smallest number of training parameters among previous models. Specifically, our ELLA$_{SDv1.5}$ and ELLA$_{SDXL}$ models require only 0.06B and 0.47B training parameters, respectively. Despite this, they demonstrate superior performance over other models across the evaluation metrics, with their performance only slightly trailing behind that of DALL-E 3. To better inspect the performance of different models, we illustrate the qualitative results on PartiPrompts in Fig. 4. We also conduct experiments to incrementally increase the complexity of the prompt. As illustrated in Fig. 6, both SDXL and DALL-E 3 exhibit difficulties in managing dense prompts, often conflating multiple objects and missing fine-grained details. In contrast, our ELLA model demonstrates a superior comprehension of the given prompt, generating images that capture more precise details.

User Study. While quantitative results can partially represent a model's performance, they may not be sufficiently comprehensive and accurate. Consequently, we conducted a user study to enrich the evaluation of the model. We employ image generation results of our DPG-Bench to examine the text-image alignment and aesthetic quality of CLIP-based SDXL, T5-based PixArt-α and our ELLA$_{SDXL}$. For each prompt with 4 images generated by distinct models, we enlist 20 unique users to rank images based on semantic alignment and aesthetic

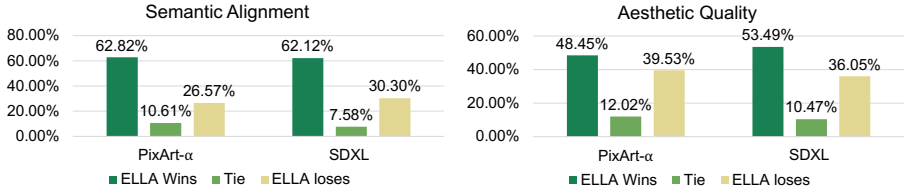

Fig. 5: The results of user study. The bar chart demonstrates that our model surpasses existing open-source models in terms of text-image alignment capabilities while maintaining a comparable aesthetic quality.

Fig. 6: The comparison between SDXL, ELLA$_{SDXL}$, and DALL-E 3 reveals their performance across varying levels of prompt complexity. Prompts range from simple to intricate from top to bottom. The results demonstrate that our model is capable of following both simple and complex prompts and generating fine-grained detail.

quality. The user study results are reported in Fig. 5. We observe that human preferences align with our evaluation outcomes on DPG-Bench in Tab. 4, thereby attesting to the reliability of DPG-Bench. Across the board, our ELLA$_{SDXL}$ outperforms the current SOTA open-source models in terms of text-image alignment, while maintaining an aesthetic quality comparable to that of SDXL.

Compatibility with Downstream Tools. As we freeze the original diffusion model in the training stage, ELLA can be seamlessly integrated into the downstream tools of Stable Diffusion. We equip six widely used community models from CivitAI [1] with ELLA to further improve their prompt following ability. As shown in Fig. 7, community models (e.g., LoRA [26], ControlNet [61]) can significantly benefit from ELLA, improving the semantic matching of generated images while maintaining a similar style to the original.

5.3 Ablation Study

We perform ablation experiments to inspect the effect of different LLM selections and alternative architecture designs. All experiments in the ablation study are conducted based on SD v1.5. It is noted that with limited computational re-

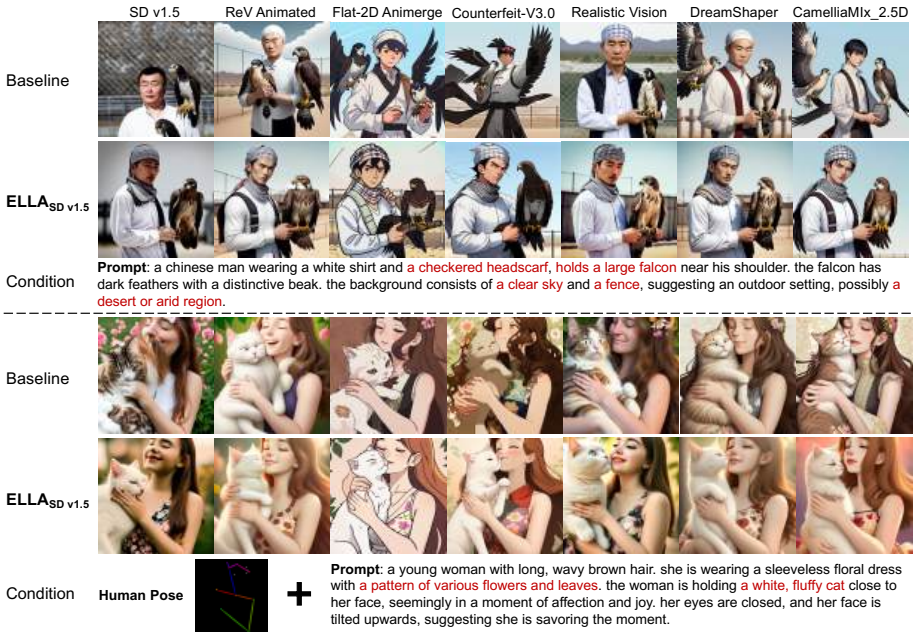

Fig. 7: Qualitative results about ELLA$_{SDv1.5}$ with personalized models. We selected representative personalized models from CivitAI [1], equipping them with ELLA$_{SDv1.5}$ to improve their prompt following ability.

sources, we train for 140,000 optimization steps (about 1 epoch) in the ablation study and 280,000 steps in the main experiments.

LLM Selection. Considering that the structure and number of parameters of different Large Language Models may result in varying capabilities in semantic understanding, we conduct experiments with 1B T5-XL encoder, 1.1B TinyLlama, and 13B LLaMA-2 on the SD v1.5. We employ the same semantic alignment connector with 6 blocks of resampler. Tab. 5 reports the corresponding performance on a subset of T2I-CompBench [27] and DPG-Bench, which assess short and complex prompts understanding respectively. Our ELLA, when equipped with various Large Language Models, consistently demonstrates superior performance over SD v1.5, which is based on CLIP. We observe that

Table 5: Ablation study on LLM selection based on SD v1.5. CLIP represents the original SD v1.5. For LLM-based text encoder, resampler with 6 blocks is applied.

Text Encoder	Attribute Binding			DPG-Bench↑
	Color↑	Shape↑	Texture↑	
CLIP	0.3750	0.3576	0.4156	63.18
TinyLlama	0.4168	0.3922	0.4639	70.27
LLaMA-2	0.4468	0.3983	0.5137	72.05
T5-XL	0.5570	0.4522	0.5195	71.70

Table 6: Ablation study on module design based on TinyLlama and SD v1.5. The last row represents our TSC design.

Module Arch.	Norm	Timestep Aware	Trainable Params	Attribute Binding			DPG-Bench↑
				Color↑	Shape↑	Texture↑	
MLP	LN	✗	2.16M	0.3262	0.3198	0.3957	62.55
Resampler (1 block)	LN	✗	8.71M	0.3569	0.3343	0.4124	66.39
Resampler (6 blocks)	LN	✗	44.16M	0.4168	0.3922	0.4639	70.27
Resampler (6 blocks)	AdaLN-Zero	✓	73.91M	0.4774	0.3810	0.4964	70.43
Resampler (6 blocks)	AdaLN	✓	66.82M	0.5014	0.4253	0.5175	72.91

for decoder-only models, LLaMA-2 with 13B parameters performs better than 1.1B TinyLlama on both short and complex prompts. Compared to decoder-only LLMs, the 1.2B T5-XL encoder shows significant advantages in short prompts interpretation while falling short of LLaMA-2 13B in comprehending complex text. We suspect that encoder models with bidirectional attention capabilities may capture richer text features, thereby providing more effective conditioning for image generation. However, the model scale of T5-XL may pose a constraint on its ability to understand intricate texts.

Module Network Design. We explore various network designs for our ELLA module to examine the effectiveness of our chosen network architecture. To this end, we conduct experiments on MLP, resampler, and resampler with timestep using AdaLN-Zero and AdaLN. All experiments are conducted with TinyLlama. Tab. 6 compares the performance of different network designs. It is observed that at a similar model scale, the transformer-based module is more effective in transferring the capabilities of language models to diffusion models than MLP. In addition, transformer blocks are more flexible for scaling up than MLP, which facilitates the upscaling of the module. We also conduct a comparative analysis of the scaled-up resampler with 6 blocks, incorporating timestep with both AdaLN-Zero and AdaLN. The latter configuration represents our final design for TSC, which performs the best on evaluation metrics. Although with more trainable parameters, experimental results illustrate that AdaLN-Zero underperforms AdaLN in our situation. AdaLN-Zero initializes each resampler block as the identity function, potentially weakening the contribution of LLM features to the final condition features.

To analyze the extracted timestep-dependant semantic features of TSC, we visualize the relative variation in attention scores between text tokens and learnable queries at different denoising timesteps in Fig. 8. Across the diffusion timestep, for each text token in the prompt, we calculate the attention and normalize it by the maximum attention score of all timesteps. Each column corresponds to a single token, demonstrating the temporal evolution of the significance attributed to each token. It is observed that the attention values of words corresponding to the primary color (i.e., blue and red) and layout (i.e., standing next to) of the image are more pronounced at higher noise levels, during which

Fig. 8: Visualization of the relative variation in attention scores between the text tokens feature and learnable queries, as influenced by the timestep.

the fundamental image formation transpires. In contrast, at lower noise levels, diffusion models generally predict high-frequency content, wherein the attention values of pertinent words (i.e., painting) that describe the image style become more prominent. This observation attests to the effectiveness of our proposed TSC in extracting semantic features from LLM across sampling timesteps. Furthermore, we note that the attention scores of the main entities, such as the cow and the tree, remain consistently strong throughout the denoising process, suggesting that diffusion models constantly focus on the primary entities during image construction.

6 Conclusion and Limitation

This paper introduces ELLA, a method that equips current text-to-image diffusion models with state-of-the-art Large Language Models without the training of LLM and U-Net. We design a lightweight and adaptive Timestep-Aware Semantic Connector (TSC) to effectively condition the image generation process with comprehensive prompts understanding from LLM. With our ELLA, the diffusion model can generate high-fidelity and accurate images according to the given long prompts with dense information. In addition, we introduce a benchmark, rewritten from various data sources, to better evaluate the model performance on long dense prompts automatically. Experimental results demonstrate that ELLA outperforms current state-of-the-art text-to-image models in prompt following ability and can be easily incorporated with community models and downstream tools. With enhanced text-image alignment, our approach sheds light on image editing in future work. Furthermore, we also plan to investigate the integration of MLLM with diffusion models, enabling the utilization of interleaved image-text input as a conditional component in the image generation process. In terms of limitations, our training captions are synthesized by MLLM, which are sensitive to the entity, color, and texture, but are usually unreliable to the shape and the spatial relationship. On the other hand, the aesthetic quality upper bound of generated images may be limited by the frozen U-Net in our approach. These can be further addressed in the future.

Acknowledgements

We thank Zebiao Huang, Huazheng Qin, Lei Huang, Zhengnan Lu, and Chen Fu for their support on data construction. We thank Fangfang Cao and Gongfan Cheng for their input on evaluation. We thank Keyi Shen and Dongcheng Xu for their support in accelerating training.

References

1. Civitai. https://civitai.com/ 11, 12
2. Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor 6
3. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 5
4. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 3, 5
5. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022) 3, 4
6. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. arXiv preprint arXiv:2302.08113 (2023) 4
7. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 1, 2, 4
8. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009) 6
9. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022) 6
10. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023) 5
11. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023) 4, 9
12. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 2, 4, 9
13. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024) 4, 5
14. Cho, J., Hu, Y., Baldridge, J., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In: ICLR (2024) 3, 7
15. Cho, J., Zala, A., Bansal, M.: Visual programming for text-to-image generation and evaluation. arXiv preprint arXiv:2305.15328 (2023) 5
16. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11472–11481 (2022) 3
17. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023) 4
18. Fang, G., Jiang, Z., Han, J., Lu, G., Xu, H., Liang, X.: Boosting text-to-image diffusion models with fine-grained semantic rewards. arXiv preprint arXiv:2305.19599 (2023) 5

19. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032 (2022) 9
20. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PUIqjT4rzq7 4, 5
21. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems **36** (2024) 5
22. Feng, Y., Gong, B., Chen, D., Shen, Y., Liu, Y., Zhou, J.: Ranni: Taming text-to-image diffusion for accurate instruction following. arXiv preprint arXiv:2311.17002 (2023) 5
23. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 5
24. Hatamizadeh, A., Song, J., Liu, G., Kautz, J., Vahdat, A.: Diffit: Diffusion vision transformers for image generation. arXiv preprint arXiv:2312.02139 (2023) 3, 6
25. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 5
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 3, 11
27. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 3, 5, 6, 9, 12
28. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: ICCV (2023) 4
29. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. arXiv preprint arXiv:2205.12005 (2022) 8, 10
30. Li, D., Kamko, A., Sabet, A., Akhgari, E., Xu, L., Doshi, S.: Playground v2. https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic 8
31. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Divide & bind your attention for improved generative semantic nursing. arXiv preprint arXiv:2307.10864 (2023) 4
32. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023) 5
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 7
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 5
35. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022) 4, 9
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 9

37. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 4
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 3, 6
39. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018) 3
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 1, 2, 4, 8, 9
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 4
42. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020) 2, 4, 5, 8
43. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 1, 2, 4
44. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36** (2024) 4
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 2, 4, 8, 9
46. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 3
47. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 1, 2, 4
48. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 6
49. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 7
50. Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.C., Alon, D., Herrmann, C., van Steenkiste, S., Krishna, R., et al.: Dreamsync: Aligning text-to-image generation with image understanding feedback. arXiv preprint arXiv:2311.17946 (2023) 5

51. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in Neural Information Processing Systems **36** (2024) 6
52. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 2, 4, 5
53. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models (2023) 6
54. Wang, Z., Xie, E., Li, A., Wang, Z., Liu, X., Li, Z.: Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation. arXiv preprint arXiv:2401.15688 (2024) 5
55. Wu, W., Li, Z., He, Y., Shou, M.Z., Shen, C., Cheng, L., Li, Y., Gao, T., Zhang, D., Wang, Z.: Paragraph-to-image generation with information-enriched diffusion model (2023) 2, 3, 4
56. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023) 4
57. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 5
58. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. arXiv preprint arXiv:2401.11708 (2024) 5
59. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 3
60. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022) 2, 3, 6, 7
61. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 3, 11
62. Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: An open-source small language model (2024) 4, 5
63. Zhong, S., Huang, Z., Wen, W., Qin, J., Lin, L.: Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 567–578 (2023) 5