Diffusion Beats Autoregressive: An Evaluation of Compositional Generation in Text-to-Image Models

Arash Marioriyad
Department of Computer Engineering
Sharif University of Technology
arash.marioriyad98@sharif.edu

Parham Rezaei
Department of Computer Engineering
Sharif University of Technology
parham.rezaei@sharif.edu

Mahdieh Soleymani Baghshah
Department of Computer Engineering
Sharif University of Technology
soleymani@sharif.edu

Mohammad Hossein Rohban
Department of Computer Engineering
Sharif University of Technology
rohban@sharif.edu

Abstract

Text-to-image (T2I) generative models, such as Stable Diffusion and DALL-E, have shown remarkable proficiency in producing high-quality, realistic, and natural images from textual descriptions. However, these models sometimes fail to accurately capture all the details specified in the input prompts, particularly concerning entities, attributes, and spatial relationships. This issue becomes more pronounced when the prompt contains novel or complex compositions, leading to what are known as compositional generation failure modes. Recently, a new open-source diffusion-based T2I model, FLUX, has been introduced, demonstrating strong performance in high-quality image generation. Additionally, autoregressive T2I models like LlamaGen have claimed competitive visual quality performance compared to diffusion-based models. In this study, we evaluate the compositional generation capabilities of these newly introduced models against established models using the T2I-CompBench benchmark. Our findings reveal that LlamaGen, as a vanilla autoregressive model, is not yet on par with state-of-the-art diffusion models for compositional generation tasks under the same criteria, such as model size and inference time. On the other hand, the open-source diffusion-based model FLUX exhibits compositional generation capabilities comparable to the state-of-the-art closed-source model DALL-E3.

1 Introduction

Recent advancements in computational resources and data scaling have led to the development of substantial text-to-image (T2I) models, from diffusion-based [16] models such as Stable Diffusion [37, 31] and DALL-E [34, 33, 4] to autoregressive-based ones such as LlamaGen [41], which are capable of producing high-quality and realistic images from textual prompts. Despite these advancements, these models occasionally face difficulties in generating images that fully align with the input prompts, especially when the prompts involve complex and novel combinations of entities, attributes, and spatial relationships [17, 3, 23]. This challenge, known as visual compositional generation, remains a significant issue in the field of T2I generation.

Compositional generation failure modes can be categorized into four main types: *Entity missing*, *incorrect attribute binding*, *incorrect spatial relationship*, and *incorrect numeracy*. Entity missing [48, 6, 2, 40, 46] is a key failure mode in T2I models, where the model omits one or more entities described in the input prompt, particularly in complex scenes involving multiple entities. Moreover,

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

incorrect attribute binding [36, 13, 25, 44] occurs when an attribute, such as color, shape, size, or texture, is not faithfully bound or associated with the corresponding entity. Furthermore, incorrect spatial relationship [14, 5] is related to the scenario where the T2I model fails to accurately capture the relative positions, orientations, or interactions between entities, resulting in a misrepresentation of the spatial arrangement described in the prompt. Finally, incorrect numeracy [47, 18] occurs when the model can not accurately represent the number of entities described in the input prompt, which reflects the model's limited reasoning abilities, as it struggles to maintain numerical consistency in complex scenes.

Several studies have explored the compositional generation capabilities of both diffusion [28] and autoregressive models [35, 10]. However, to the best of our knowledge, the field lacks a comprehensive comparison of these two generative approaches in the context of visual compositional generation from textual prompts. This study extensively evaluates the compositional generation abilities of nine state-of-the-art T2I models, including seven diffusion-based and two autoregressive-based models, using the established benchmark, T2I-CompBench [17].

Our results indicate that the vanilla autoregressive-based T2I model, LlamaGen [41], underperforms in all compositional generation assessments compared to SD-v1.4, the diffusion-based model most similar to LlamaGen in terms of model size (number of parameters) and inference time. This finding may suggest that adhering solely to the next-token prediction paradigm, without incorporating additional inductive biases, is insufficient to match the performance of diffusion-based approaches in compositional generation. Furthermore, an evaluation of the newly introduced open-source diffusion-based model, FLUX [21], demonstrates that it performs competitively with the state-of-the-art closed-source T2I model, DALL-E3 [4].

2 Text-to-image Models

This assessment evaluated nine famous T2I backbones, comprising seven diffusion-based and two autoregressive-based models (Table 1). These models can be categorized into five distinct families.

Stable Diffusion: Stable Diffusion models are among the most prominent open-source T2I models, utilizing a latent diffusion framework combined with an attention mechanism to process textual prompts. Specifically, the process begins with pure noise sampled from a Gaussian distribution as the initial latent code. The model then iteratively refines the latent code at each denoising step using a U-Net [38] architecture, which incorporates cross-attention layers to align the image generation process with the textual embedding obtained from a CLIP-based [32] model. After a predefined number of denoising steps, the final refined latent code is passed through a pre-trained image decoder to generate the final image. Through this work, we employed SD-v1.4 , SD-v2 [37], and SD-XL [31] from the Stable Diffusion family.

DALL-E: DALL-E models are a family of closed-source diffusion-based T2I models developed and maintained by OpenAI [29]. While DALL-E1 [34] utilizes a discrete variational auto-encoder [43] model to generate image tokens from textual tokens, DALL-E2 [33] first uses a pre-trained CLIP-based model to prepare the text embeddings from the input prompt, which is then fed to a diffusion or autoregressive model to produce an image embedding. Finally, a diffusion decoder conditioned on the obtained embedding produces the final image. To further improve the prompt following abilities and image quality, DALL-E3 [4] adopts a recaptioning process of the training dataset, which is then used as the new training data for the T2I model.

Pixart-α: Pixart-α [7] utilizes the Diffusion Transformer (DiT) [30] as its core architecture, prioritizing rapid and cost-effective training compared to Stable Diffusion models. Specifically, Pixart-α employs a three-stage training strategy, along with a recaptioning process for the training data. In the first stage, known as pixel dependency learning, Pixart-α benefits from parameter initialization derived from an ImageNet-pretrained model and a class-guided approach to image generation. This phase focuses on generating semantically coherent pixels during a relatively inexpensive training process. The second stage involves text-image alignment learning, where Pixart-α constructs a dataset with precise text-image pairs that exhibit high concept density, utilizing the advanced vision-language model LLaVA [27] applied to the SA-1B dataset [19]. In the final stage, the model is fine-tuned with high-quality aesthetic data to enhance its capability for high-resolution image generation.

FLUX: FLUX family models [21] are newly introduced T2I models developed by Black Forest Lab [20]. To the best of our knowledge, no formal technical report is available about this model. However, based on the available implementation details, the FLUX family models employ a hybrid architecture that integrates multi-modal [11] and parallel [9] diffusion transformer [30] blocks, operating within a flow-matching [26] framework. Additionally, FLUX utilizes rotary positional embeddings [39] and parallel attention layers [9]. The FLUX models are available in three versions: Pro, Schnell, and Dev, with the latter two being utilized in our evaluations.

Autoregressive: Vanilla Autoregressive T2I models such as LlamaGen [41] employ the next-token prediction paradigm, commonly seen in large language models (LLMs), for image generation. Particularly, LlamaGen utilizes the Llama [42] architecture for pixel generation and a quantized-autoencoder [12] framework for image tokenization. LlamaGen is introduced in two variants: class-conditioned and text-conditioned, with the latter being used for our compositional generation evaluation. For the text-conditioned variant, the model follows a two-stage training strategy. The first phase involves training on a 50-million subset of the LAION-COCO dataset [22]at a resolution of 256×256, followed by a fine-tuning phase on 10 million internally curated high-aesthetic-quality images at a resolution of 512×512, as the second phase. LlamaGen does not incorporate a diffusion-based process, positioning it as a vanilla autoregressive T2I model without additional inductive biases.

Table 1: Detailed Information on State-of-the-art Text-to-image Models

Model	Release Date	#Parameters	Resolution	Training Data	Text Encoder
SD-v1.4	Aug 2022	860×10^6	512×512	LAION-5B	ViT-L/14 CLIP
SD-v2	Nov 2022	860×10^6	768×768	LAION-5B	ViT-H/14 CLIP
SD-XL	July 2023	3.5×10^9	1024×1024	LAION-5B	OpenCLIP-ViT/G
DALL-E3	Nov 2023	-	1024×1024	-	CLIP-based
Pixart-α	Sep 2023	600×10^6	1024×1024	SA-1B	T5
FLUX-Dev	Aug 2024	-	512×512	-	T5 & CLIP-based
FLUX-Schnell	Aug 2024	-	1024×1024	-	T5 & CLIP-based
LlamaGen-Stage1	June 2024	775×10^6	256×256	LAION-COCO	T5
LlamaGen-Stage2	June 2024	775×10^6	512×512	-	T5

3 T2I-CompBench Benchmark

3.1 Evaluation Datasets

The T2I-CompBench dataset [17] evaluates four main aspects of compositional generation capabilities: attribute binding, object relationships, numeracy, and complex compositions.

Attribute Binding: This section is divided into three categories—color, shape, and texture—each comprising 300 validation prompts.

Object Relationships: This part is further split into spatial and non-spatial relationships. The spatial relationships involve two sets of prompts, 2D and 3D, with 300 validation prompts each. The non-spatial relationships focus on interactions between objects and also contain 300 validation prompts.

Numeracy: This section includes 300 validation prompts to assess the ability of the T2I model to understand and reason about numerical concepts.

Complex Compositions: This part presents 300 validation prompts that feature more natural and challenging combinations of objects, attributes, and relationships.

3.2 Evaluation Metrics

T2I-CompBench [17] employs a visual question answering (VQA) model, called BLIP-VQA [24], to evaluate the attribute binding capabilities of T2I models. For assessing spatial relationships and numeracy, the framework uses an object detector, UniDet [49], to estimate the relational positions and the number of objects in the generated images. For non-spatial relationships and complex compositions, T2I-CompBench utilizes several evaluation metrics: CLIP similarity score [15], which measures cosine similarity between the embeddings of the prompt and the generated image using a CLIP-based [32] model; multi-modal evaluation via a GPT-based model [1]; chain-of-thought prompting [45] using ShareGPT-4v [8]; and finally, the 3-in-1 evaluation metric, which integrates CLIP similarity, BLIP-VQA, and UniDet scores.

4 Results

Table 2 presents the results of attribute binding, spatial relationship, and numeracy assessments for nine state-of-the-art T2I models. DALL-E 3 and FLUX-based models demonstrate competitive performance across all aspects of compositional generation, consistently ranking at the top. Following these models, the Pixart model outperforms SD-XL in most evaluations. In contrast, the vanilla autoregressive model, LlamaGen, underperforms even when compared to the weakest Stable Diffusion model, SD-v1.4, which is comparable to LlamaGen in terms of model size and inference time. Similar trends are observed in Table 3, which reports the results of non-spatial relationships and complex composition assessments.

Table 2: Quantitative Results of T2I Models on Attribute Binding, Spatial Relationship, and Numeracy

Model	Color (↑)	Shape (↑)	Texture (↑)	2D-Spatial (↑)	3D-Spatial (↑)	Numeracy (↑)
SD-v1.4	0.376	0.358	0.416	0.125	0.303	0.446
SD-v2	0.506	0.422	0.492	0.134	0.323	0.458
SD-XL	0.588	0.469	0.530	0.213	0.357	0.499
DALL-E3	**0.778**	**0.620**	**0.704**	<u>0.286</u>	0.374	0.588
Pixart-α	0.670	0.493	0.648	0.206	<u>0.390</u>	0.506
FLUX-Dev	<u>0.771</u>	0.495	0.604	0.266	0.384	**0.618**
FLUX-Schnell	0.740	<u>0.571</u>	<u>0.685</u>	**0.292**	**0.391**	<u>0.606</u>
LlamaGen-Stage1	0.271	0.391	0.492	0.084	0.227	0.357
LlamaGen-Stage2	0.285	0.329	0.373	0.119	0.155	0.265

Table 3: Quantitative Results of T2I Models on Non-spatial Relationship and Complex Compositions

	Non-spatial (↑)			Complex (↑)		
Models	CLIP	GPT-4v	Share-CoT	3-in-1	GPT-4v	Share-CoT
SD-v1.4	0.308	0.820	0.749	0.308	0.714	0.773
SD-XL	<u>0.312</u>	0.844	0.767	0.324	0.756	0.782
DALL-E3	0.300	**0.927**	**0.793**	**0.377**	**0.828**	**0.793**
FLUX-Dev	0.306	<u>0.874</u>	0.780	0.364	0.794	<u>0.791</u>
FLUX-Schnell	**0.313**	0.872	<u>0.784</u>	<u>0.368</u>	<u>0.823</u>	**0.793**
LlamaGen-Stage1	0.305	0.788	0.783	0.283	0.584	0.769
LlamaGen-Stage2	0.272	0.669	0.763	0.255	0.562	0.765

5 Discussion

These findings indicate that the pure next-token prediction paradigm fails to compete effectively with diffusion-based models of similar size in the absence of inductive biases tailored to the visual generation domain. Notably, While the class-conditioned version of LlamaGen exhibits competitive performance in terms of image quality and naturalness compared to diffusion models, the weaker performance of the text-conditioned version suggests that autoregressive models may face more significant challenges in capturing complex conditions. Furthermore, considering the critical role of tokenization in autoregressive models, selecting an appropriate image tokenizer with suitable granularity may enhance LlamaGen's compositional generation capabilities. Ultimately, the inductive bias of the mask image modeling or next-token prediction paradigm may not be sufficient for generating images that are fully aligned with the textual prompts.

References

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

[2] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023.

[3] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.

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

[5] Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. Getting it right: Improving spatial consistency in text-to-image models. *arXiv preprint arXiv:2404.01197*, 2024.

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

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

[8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[9] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[10] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[13] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *ArXiv*, abs/2212.10015, 2022.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

[17] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[18] Wonjun Kang, Kevin Galim, and Hyung Il Koo. Counting guidance for high fidelity text-to-image synthesis. *arXiv preprint arXiv:2306.17567*, 2023.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[20] Black Forest Lab. `https://www.blackforestlab.com/blog`, 2024. Accessed: 2024-09.

[21] Black Forest Lab. Flux: A diffusion-based text-to-image (t2i) model. `https://github.com/blackforestlab/flux`, 2024. Accessed: 2024-09.

[22] LAION. Laion-coco 600m. `https://laion.ai/blog/laion-coco`, 2022.

[23] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[25] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *34th British Machine Vision Conference 2023, BMVC 2023*, 2023.

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[28] Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024.

[29] OpenAI. `https://www.openai.com`, 2024. Accessed: 2024-09.

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

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

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

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

[35] Rahul Ramesh, Ekdeep Singh Lubana, Mikail Khona, Robert P. Dick, and Hidenori Tanaka. Compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks. In *Forty-first International Conference on Machine Learning*, 2024.

[36] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[39] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[40] Kota Sueyoshi and Takashi Matsubara. Predicated diffusion: Predicate logic-based attention guidance for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8651–8660, 2024.

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

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[43] Arash Vahdat, Evgeny Andriyash, and William Macready. Dvae#: Discrete variational autoencoders with relaxed boltzmann priors. *Advances in Neural Information Processing Systems*, 31, 2018.

[44] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5544–5552, 2024.

[45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[46] Yihang Wu, Xiao Cao, Kaixin Li, Zitan Chen, Haonan Wang, Lei Meng, and Zhiyong Huang. Towards better text-to-image generation alignment via attention modulation. *arXiv preprint arXiv:2404.13899*, 2024.

[47] Oz Zafar, Lior Wolf, and Idan Schwartz. Iterative object count optimization for text-to-image diffusion models. *arXiv preprint arXiv:2408.11721*, 2024.

[48] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Tiviatis Sim, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. *arXiv preprint arXiv:2403.06381*, 2024.

[49] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7571–7580, 2022.