ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890155

Semantic Attention and LLM-based Layout Guidance for Text-to-Image Generation

Yuxiang Song[1], Zhaoguang Long[1], Man Lan[1,*], Changzhi Sun[2], Aimin Zhou[1], Yuefeng Chen[3], Hao Yuan[3], and Fei Cao[3]
[1] *School of Computer Science and Technology, East China Normal University, Shanghai, China*
[2] *Institute of Artificial Intelligence (TeleAI), China Telecom*
[3] *Shanghai Transsion Co., Ltd., Shanghai, China*
yxsong@stu.ecnu.edu.cn, mlan@cs.ecnu.edu.cn

***Abstract*—Diffusion models have substantially advanced text-to-image generation, achieving remarkable performance in creating high-quality images from textual prompts. However, they often struggle with accurately generating images representing spatial locations described or implied in the prompts. To address this, we introduce SALT, a training-free method leveraging semantic attention and layout guidance from Large Language Models (LLMs) for text-to-image generation. This method effectively guides both cross-attention and self-attention layers within diffusion models, steering generation toward the direction of high-attention values provided by the layout guidance. During the denoising process of the diffusion model, image features in the latent space are iteratively refined based on the loss function calculated from the desired attention maps. Our approach has been executed on two benchmarks, providing detailed qualitative examples and comprehensive quantitative analyses. Results demonstrate that SALT outperforms existing training-free methods in controlling object layouts and generating attributes.[1]**

***Index Terms*—diffusion models, text-to-image generation, attention mechanism, training-free method**

I. Introduction

Recent advancements in Text-to-Image (T2I) generation models based on diffusion techniques represent a significant stride in cross-modal learning. Modern impactive generative diffusion models, (e.g., DALL-E [1], Imagen [6], and Stable Diffusion [8]), have unlocked limitless possibilities for Artificial Intelligence Generated Content (AIGC) processes. These models enable the synthesis of diverse and realistic images with flexible editing capabilities. Despite the substantial progress in T2I model development, they often have limitations in following strictly textual descriptions, frequently exhibiting significant hallucination issues. An obvious limitation of these models is their inability to generate images that accurately depict the spatial locations. This challenge predominantly arises from the limitations of the CLIP [2] text encoder used in the diffusion models, which struggles to interpret complex spatial descriptions.

Efforts in the field have recently concentrated on enhancing the controllability of pre-trained diffusion models, particularly within the Layout-to-Image (L2I) Generation. Researchers [15]–[17] have proposed layout-conditioned diffusion models to tackle this challenge by training or fine-tuning text-to-image diffusion models with additional network layers that provide spatial control, training adjustments, and enriched datasets. [16] introduced LayoutDiffusion, treating each image block as a distinct object within the layout and facilitating the generation of high-quality and diverse images while enabling precise control over the positioning and sizing of multiple objects. [15] employs gated self-attention layers designed to add extra inputs, such as bounding boxes, enhancing spatial manipulation capabilities. [17] incorporates regional tags into textual prompts, achieving layout control through encoding the fixed prompts.

However, the process of training or fine-tuning can be computationally expensive. Moreover, models need to be retrained for each new base model. In contrast, training-free methods have emerged as an alternative, manipulating cross-attention layers to guide image generation without additional training. These methods are typically categorized into two types [14]: forward guidance, which directly imposes the attention layers to align activations with the desired patterns, and backward guidance, which adjusts latent variables through gradient-based updates.

[11] extracts cross-attention maps for each text token and modifies these values to control the generation of images. [13] introduces a method to adjust cross-attention maps during the denoising period, strengthening the attention to the selected tokens in the textual prompts. [14] also manipulates the cross-attention layers, guiding the reconstruction within the user-specified layout using two distinct strategies. While these methods meticulously design the required conditions to manipulate cross-attention maps, enhancing interactions between textual and visual information, they often overlook the manipulation of self-attention layers, which are crucial for handling communication between different object features. Furthermore, the aforementioned methods require the troublesome manual design of certain conditions during implementation. Recently, research leveraging LLMs [24]–[26] for layout generation has begun to emerge [9], [10], which utilizes LLMs to infer spatial concepts under textual conditions. [18] introduces a plug-and-play approach, which employs attention refocusing to handle foreground and background regions. The above method has already proven that adding layout guidance during the attention map generation process is crucial for text-to-image generation. While our technical approach is similar, in contrast, our method focuses on directly enhancing the high-attention value regions in both attention maps during each denoising step, achieving good results in a straightforward way.

Building upon these findings, we propose an innovative method that utilizes LLMs to generate bounding boxes as layout guidance for objects mentioned in textual prompts. This method integrates attention-guided optimization with layout guidance during the denoising process in text-to-image generation. The combined approach ensures images comply with spatial constraints while maintaining high visual fidelity.

In summary, the contributions of our paper are as follows:

- We introduce a novel training-free method that merges LLM-generated bounding boxes with attention-guided optimization in text-to-image generation without the manually designed conditions.
- By manipulating both cross-attention and self-attention layers, we direct heightened attention to specified areas, promoting objects

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology(72192820 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05). (* Corresponding author: Man Lan)

[1]Our code and dataset are released at https://github.com/cubenlp/SALT

are generated and positioned as anticipated.
- Compared to other training-free methods that manipulate attention, our approach achieves better generation results, as demonstrated on the Drawbench and HRS datasets.

II. Methodology

A. Preliminaries

*1) **Diffusion Models**:* Diffusion models are a class of generative models based on probabilistic principles, comprising both a forward diffusion process and a reverse diffusion process. In text-to-image generation, the reverse denoising process plays a pivotal role. The model learns to progressively denoise the input by accurately predicting the noise added during the forward process. This iterative denoising process is crucial to the effectiveness of diffusion models.

Stable Diffusion is a state-of-the-art image generation model based on the diffusion framework. Unlike earlier models that manipulate image data at the pixel level, it uses a pre-trained autoencoder to operate within a compressed latent space, providing a more efficient way to generate higher-quality images. The model is conditioned on textual inputs, which are encoded by a pre-trained CLIP text encoder. Given a conditioning prompt p, the corresponding conditioning vector $c(p)$ is integrated into the diffusion process.

The model's training objective is to minimize the following loss function: $\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x)}\left[\|\epsilon_t - \epsilon_\theta(z_t, t, c(p))\|_2^2\right]$, where the encoder $\mathcal{E}$ maps an image x into a latent space, producing the encoded latent representation z. At each timestep t, noise of different levels is added to z, resulting in the latent representation z_t. The diffusion model ϵ_θ, such as the U-Net architecture with a scheduler, is conditioned on the text embedding $c(p)$ to predict the noise added to z_t, with θ represents learnable parameters, while ϵ_t denotes the actual Gaussian noise associated with z_t.

*2) **Attention Layers**:* Attention layers [3], consisting of both self-attention and cross-attention layers, are critical components of the denoising U-Net architecture in Stable Diffusion, operating at resolutions of 64, 32, 16, and 8. Self-attention layers facilitate the utilization of global information, allowing for the synthesis of globally coherent structures by linking disparate regions of an image. Cross-attention layers serve as bridges between textual and image modalities, typically employing a pre-trained CLIP encoder to process textual prompts, resulting in text embedding features. Keys and values are derived from text embedding features through linear mappings. The attention matrix is defined as follows:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q and K represent the queries and keys within the attention layer of the transformer block, while V denotes this layer's value, d_k serves as a scaling factor for the dimension of the key.

B. LLM-based Layout Generation

LLMs exhibit advanced spatial reasoning capabilities, enabling them to accurately model the spatial positions of objects. We instruct the LLM to generate bounding boxes to provide essential spatial structure and layout guidance for text-to-image generation, as depicted in Figure 2. The process is divided into three primary steps: (1) Generating bounding boxes with coordinates formatted as (x_1, y_1, x_2, y_2), where the x and y dimensions are normalized to the $[0, 1]$ range; (2) Associating each bounding box with a specific object mentioned in the text prompt; (3) Supplying the LLM with a manually curated example for generating layouts by employing in-context learning [23].

C. Semantic Attention and Layout Guidance

As seen in Figure 1, self-attention maps highlight areas of similar color or texture, capturing intrinsic relationships between regions of the image. In contrast, cross-attention maps serve to link specific regions of the image to tokens from the text prompt. Together, the two attention mechanisms, cross-attention, which guides the alignment between image and text, and self-attention, which focuses on intra-image coherence collaboratively contribute to the production of high-quality images that are both semantically rich and visually coherent. During the reverse diffusion process, we apply layout control as soft constraints, guiding the attention mechanism to focus more on regions within the bounding boxes during the gradient update step.

*1) **Semantic Attention Map Aggeration**:* The pre-trained CLIP text encoder is utilized to process text. It tokenizes the prompt into a sequence of tokens and transforms these tokens into a set of embeddings: $E = \{e_1, e_2, ..., e_N\}$ where $E \in R^{N\times M}$, where M is the embedding dimension, with special tokens $\langle sot\rangle$ and $\langle eot\rangle$ used to mark the start and end of the text, respectively. Each bounding box B_j associated with the phrase is defined as $B_j = (x1_j, y1_j, x2_j, y2_j)$, where the coordinates specify the left-top, and right-bottom boundaries of the j-th box. These boxes correspond to phrases $P = (p_1, p_2, ..., p_j)$ that describe the objects within. During the diffusion process, we utilize the U-Net that incorporates attention maps at a resolution of $16 \times 16 \times S$, as these maps of this resolution have been shown to contain the most semantic information [11], [13]. Here, S represents the length of text tokens in the cross-attention map and the depth of feature maps in the self-attention map. We aggregate the attention from different layers and heads at the specified resolution to obtain the aggregated cross-attention maps a_{cross} and self-attention maps a_{self}. We removed the $\langle sot\rangle$ and $\langle eot\rangle$ because they carried rich semantic and layout information. For each bounding box B_j, we define a mask M_{b_j} to isolate attention within the designated region:

$$M_{b_j} = \begin{cases} 1, & \text{if } (x1_j, y1_j, x2_j, y2_j) \in B_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This mask is then applied to the aggregated attention map, concentrating attention on the regions of interest. Subsequently, we apply Gaussian smoothing to the attention map, as proposed by [13], to ensure a uniform distribution of attention within the bounding box and to smooth any abrupt transitions. This procedure is defined as $a_{\text{cross}} = \text{GaussianSmooth}(a^j_{\text{cross}} \odot M_{b_j}, \sigma)$, where σ denotes the standard deviation. The maximum value is extracted from the cross-attention maps within each mask, with the complete set of maximum attention values defined as $\mathbf{A}^{\max}_{\text{cross}} = \{a^{\max}_{\text{cross},1}, a^{\max}_{\text{cross},2}, \ldots, a^{\max}_{\text{cross},j}\}$ where j equals the number of target regions, representing the focus of each token on specific regions during T2I cross-modal interactions.

Self-attention maps A_{self} represent the influence each pixel in the image has on and receives from other pixels, thereby capturing the internal relationships and structures within the image. Similar to the process applied to cross-attention maps, we can compute $a_{\text{self}} = \text{GaussianSmooth}(a^k_{\text{self}} \odot M_{b_j}, \sigma)$. The complete set of maximum attention values is defined as $\mathbf{A}^{\max}_{\text{self}} = \{a^{\max}_{\text{self},1}, a^{\max}_{\text{self},2}, \ldots, a^{\max}_{\text{self},k}\}$, where k equals the number of target regions multiplied by the number of pixels in the spatial dimension, indicating the model's multi-level focus on each target region under the self-attention mechanism.

*2) **Overall Loss**:* Once the final attention maps are obtained, the latent space is iteratively refined during the reverse diffusion process. At each step, the loss is computed using the maximum attention values from both the cross-attention and self-attention layers.

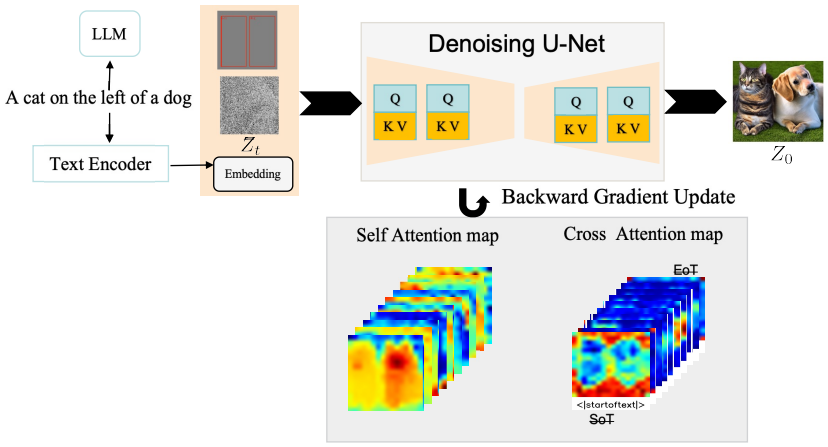

Fig. 1. Method Overview. We utilize LLMs to generate bounding boxes, extracting corresponding self-attention and cross-attention maps. During the inference process, we optimize the latent at each timestep through backpropagation, updating the latent z_t accordingly. The attention maps use brighter colors to signify higher attention scores, effectively visualizing the focus of the model.

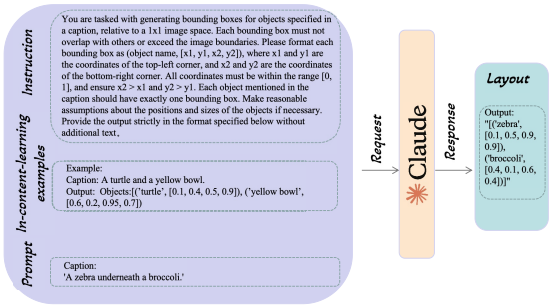

Fig. 2. The diagram of layout generation.

For the cross-attention maps, the maximum attention value for each region, $a^{\max}_{\text{cross},j}$, is used to define the first part of the loss function $\mathcal{L}_{\text{cross}}$, assuming an ideal maximum attention value of 1. Similarly, for the self-attention maps corresponding to each bounding box, the second part of the loss function $\mathcal{L}_{\text{self}}$ is defined based on the maximum attention values, $a^{\max}_{\text{self},k}$, under the same assumption.

$$\begin{aligned} \mathcal{L}_{\text{cross}} &= \max_j \left[\max(0, 1 - a^{\max}_{\text{cross},j})\right] \\ \mathcal{L}_{\text{self}} &= \max_k \left[\max(0, 1 - a^{\max}_{\text{self},k})\right] \end{aligned} \tag{3}$$

The overall loss is the sum of the cross-attention and self-attention losses: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}}$.

*3) **Backward Gradient Update**:* After computing the total loss $\mathcal{L}_{\text{total}}$, we optimize the noise vector z_t in the latent space at the current timestep:

$$z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}_{\text{total}} \tag{4}$$

Here, α_t represents the step size for gradient updates, and ∇ denotes the gradient operator. The updated latent variable z'_t, along with the necessary parameters, is then fed into the Stable Diffusion model to compute the noise vector for the next denoising step, z_{t-1}. This process is iteratively repeated throughout the early stages of the denoising process, from $t = T$ down to T_{end}, guiding the model to generate preliminary images that align accurately with the text prompts. During this phase, the expected object locations are carefully aligned with the bounding boxes.

Inadequate optimization during the early stage can result in unclear attention maps and potential object loss. To mitigate this, specific timesteps are selected for progressively refined gradient updates, achieving a balance between image clarity and data coherence. Further implementation details are provided in the experiment section. In the later stages of the denoising process, attention maps are no longer used to guide the model. Instead, the standard denoising steps of the Stable Diffusion model are executed directly, allowing the preliminary image to gradually regain detail while ensuring the primary object's position remains unchanged.

III. Experiment

A. Baselines and Settings

We compared our method against various training-free approaches designed to control the generative path during inference. These include MultiDiffusion [12], Attend-and-Excite [13], and Layout-Guidance [14]. Additionally, we compared our method with GLIGEN [15], which incorporates an additional gated self-attention layer trained on extensive datasets to learn new localized conditions. We included it in our experiments as it serves as an excellent benchmark for spatially conditioned L2I tasks. We also evaluated our approach against the Stable Diffusion model [8], trained on the LAION-5B dataset [4]. Following the basic settings in [13], we optimize the step size controlling the denoising process, with an additional scaling factor α_t. The iterative timesteps are set as $\{0, 10, 20\}$. Images were generated using the official Stable Diffusion model in a 50-step denoising process, with the maximum number of iterations for optimizing the loss set at 20. A Gaussian filter, characterized by a kernel size of 3 and a function f_σ defining the standard deviation as $\sigma = 0.5$, was used to smooth the attention maps.

B. Datasets and Metrics

To quantitatively assess our approach, we employed two established benchmarks: DrawBench [6] and HRS [5], which include prompts with carefully designed spatial relationships (e.g., above, below, left, right), encompassing a diverse range of textual expressions related to spatial positioning. The DrawBench dataset consists of 20 spatial prompts, for which we manually created labels based on object relationships. We also cleaned the HRS dataset, retaining 898 spatial relationship prompts, each labelled with objects and their relative positions. We used **accuracy** as the evaluation metric, considering generated images

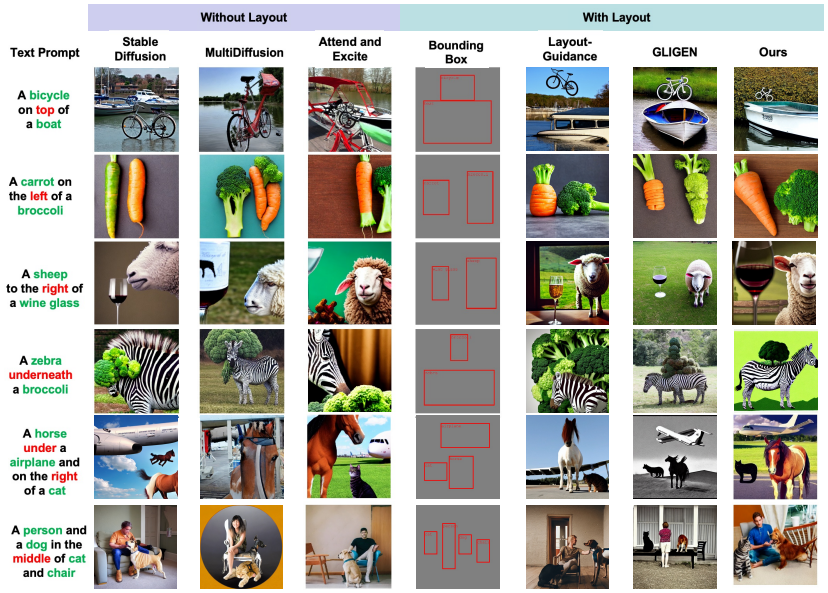

Fig. 3. Qualitative Comparison: In the textual prompts, green text denote objects, while red text highlights relationships. Methods without layout often omit objects and mostly do not follow the spatial layout described in the text (first three columns). In contrast, methods with layout, such as GLIGEN [15], which trains an additional gated self-attention layer, excel at capturing spatial attributes. Our method, however, does not involve training but uses attention and layout guidance (last column). To select the best-generated results, each method generated three images using three random seeds, from which the best were chosen.

correct when the detected objects are accurate and satisfy the specified spatial relationships. Following the HRS [5] protocol, we use the same detection setup UniDet [27] to compute metrics for generated images.

C. Experimental Results and Analysis

The quantitative results are presented in Table I, since the exact number of iterations is not specified in [18], we used three iterations as the baseline for averaging. Additionally, we randomly selected a diverse set of challenging spatial relationship prompts and manually verified them. In the DrawBench benchmark, SALT achieved the highest accuracy, scoring 55.00%, significantly outperforming other methods, including the extensively trained GLIGEN [15]. Similarly, in the HRS benchmark, our model achieved an accuracy of 18.11%, surpassing all training-free baselines.

Figure 3 illustrates the disparities in performance between methods with and without layout guidance. For instance, the Stable Diffusion [8] models often misplace objects or fail to generate them following the prompts. MultiDiffusion [12] and Attend and Excite [13] have established multi-regional generation models and enhanced text token cross-attention values respectively. While these methods show marked improvements in object attribute clarity, they still struggle with or incorrectly represent spatial relationships.

Methods incorporating layout guidance, such as GLIGEN [15], which involves training additional gated self-attention layers. It treats bounding boxes as hard constraints, requiring that objects remain confined within the designated regions as a goal. However, our method offers distinct advantages. First, it is training-free, using bounding boxes as guides to enhance attention weights within the specified areas without strictly confining object generation to these regions. This provides greater flexibility in generation. Second, compared to Layout-Guidance [14] that relies on manually designed object and positional information, our model not only aligns more closely with textual prompts but also implements a fully automated pipeline and focuses on attention guidance. Based on our analysis, while some objects may not be entirely within the bounding boxes, the majority of high-attention values are typically concentrated within or near the boxes. This enhances alignment with textual prompts and maintains flexibility in image generation.

TABLE I
QUANTITATIVE EVALUATION ON THE DRAWBENCH AND HRS BENCHMARK. THE RESULTS OF OTHER METHODS ARE FROM [18].

Method	DrawBench	HRS
	Accurancy(%) ↑	
Stable Diffusion [8]	12.50	8.48
Attend-and-Excite [13]	20.50	9.98
Layout-Guidance [14]	36.50	16.47
MultiDiffusion [12]	38.00	14.27
GLIGEN [15]	48.00	**30.74**
SALT	**55.00**	18.11

IV. CONCLUSION

In this paper, we introduced SALT, a training-free approach that leverages semantic attention and layout guidance from LLMs to enhance text-to-image generation. Without requiring additional training or fine-tuning, SALT enables users to create images that align more effectively with their needs, reducing the reliance on extensive prompt engineering. Our comparative analyses on benchmark datasets demonstrate that SALT outperforms traditional methods reliant on manually designed object indices and positional information, showcasing its ability to generate images that closely match textual prompts. However, the method has some limitations, such as the performance on multi-object prompts remains suboptimal. We will delve deeper into improving the model's capabilities in handling more complex T2I generation tasks, aiming to expand the applicability and robustness of this method in future work.

REFERENCES

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[3] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[4] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[5] E. M. Bakr, P. Sun, X. Shen, F. F. Khan, L. E. Li, and M. Elhoseiny, "Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 041–20 053.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[7] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7701–7711.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] W. Feng, W. Zhu, T.-j. Fu, V. Jampani, A. Akula, X. He, S. Basu, X. E. Wang, and W. Y. Wang, "Layoutgpt: Compositional visual planning and generation with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] L. Lian, B. Li, A. Yala, and T. Darrell, "Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models," *arXiv preprint arXiv:2305.13655*, 2023.

[11] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control.(2022)," *URL https://arxiv. org/abs/2208.01626*, 2022.

[12] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," 2023.

[13] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[14] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5343–5353.

[15] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.

[16] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 490–22 499.

[17] Z. Yang, J. Wang, Z. Gan, L. Li, K. Lin, C. Wu, N. Duan, Z. Liu, C. Liu, M. Zeng *et al.*, "Reco: Region-controlled text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 246–14 255.

[18] Q. Phung, S. Ge, and J.-B. Huang, "Grounded text-to-image synthesis with attention refocusing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7932–7942.

[19] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[20] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304.

[21] S. Zhong, Z. Huang, W. Wen, J. Qin, and L. Lin, "Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 567–578.

[22] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[24] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: open and efficient foundation language models. arxiv," *arXiv preprint arXiv:2302.13971*, 2023.

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

[27] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7571–7580.