This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

Compositional Text-to-Image Synthesis with Training-Free Layout-Guided Diffusion

Xiaoling Gu, Lingwei Luo, Shengqi Wu, Zizhao Wu, Zhenzhong Kuang and Zhou Yu

***Abstract*—Recent text-to-image (T2I) diffusion models have made significant strides in generating high-quality images from diverse textual prompts. Despite this progress, these models often face challenges in accurately understanding and synthesizing complex prompts, primarily due to their limited compositional capabilities. In this study, we propose a novel approach for compositional T2I synthesis using layout-guided diffusion models, which do not require additional training. Specifically, we leverage the chain-of-code prompting technique of large language models to interpret textual prompts and generate object layouts with spatial coherence. To enhance the alignment between generated images and textual descriptions, we introduce two innovative layout-guided loss functions: Patch-oriented Cross-Attention (PCA) loss and Region-oriented Cross-Attention (RCA) loss. The PCA loss emphasizes high activation values for image patches that attend to all tokens in the prompt across the layout. The RCA loss enhances the average attention within the layout, thereby increasing the accuracy of generating objects and their associated attributes within specified regions. These proposed loss functions reassign cross-attention in diffusion models during the denoising process. Our comprehensive experiments consistently demonstrate the effectiveness of our approach in improving semantic alignment between generated images and a diverse range of textual prompts, while ensuring high usability as a ready-to-use plugin. Our code is available at https://github.com/gxl-groups/Compositional-T2I.**

***Index Terms*—diffusion models, text-to-image synthesis, cross-attention**

Fig. 1. Compositional challenges encountered by SD-based models.

I. Introduction

Text-to-image (T2I) synthesis aims to produce high-quality and diverse images from free-form textual prompts [1]–[6]. Recently, diffusion models [7]–[10] have transformed the field of image synthesis. Among them, Stable Diffusion (SD) [7] has emerged as a leading open-source solution, demonstrating significant performance improvements following extensive training on billions of text-image pairs. However, creating high-quality images based on complex textual descriptions remains a formidable challenge. For instance, when dealing with textual prompts involving multiple objects linked by attributes and intricate spatial relationships, SD-based models often struggle with precisely composing these elements within an image. Fig. 1 provides an illustrative overview of the compositional challenges encountered by SD-based models, which can be broadly categorized into six distinct types: (1) *Object Omission*, where certain objects mentioned in the textual prompt are absent from the generated image; (2) *Object Leakage*, the inappropriate substitution of one object with another; (3) *Attribute Exchange*, involving the swapping of attributes between objects; (4) *Attribute Leakage*, where the attribute of one object is mistakenly observed in another; (5) *Spatial Misalignment*, where the spatial relationships between objects are inaccurately represented; and (6) *Interaction Neglect*, the neglect of interactions between objects in the generated image.

Previous research has highlighted the strong semantic associations between object layouts and content with the keys and values embedded in the cross-attention layers of SD-based models [11]. Consequently, some studies have attempted to mitigate these issues by adjusting the cross-attention maps during the sampling phase, without additional training or auxiliary models [12]–[14]. For example, Attend-and-Excite [12] directs a pre-trained diffusion model and refines cross-attention maps during sampling by employing a regularization loss to focus attention on the most overlooked subject tokens. Structured Diffusion [14] uses a parsing tree to extract the linguistic structure of textual prompts for diffusion guidance and adjusts cross-attention layers in diffusion-based T2I models. Nevertheless, these methods are still confronted with the aforementioned compositional challenges. For example, as

Xiaoling Gu, Lingwei Luo, Shengqi Wu, Zizhao Wu, Zhenzhong Kuang and Zhou Yu are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Zhejiang Province, China (e-mail: guxl@hdu.edu.cn; llingwei@hdu.edu.cn; wusq@hdu.edu.cn; wuzizhao@hdu.edu.cn; zzkuang@hdu.edu.cn; yuz@hdu.edu.cn).

Manuscript received April 19, 2021; revised August 16, 2021.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

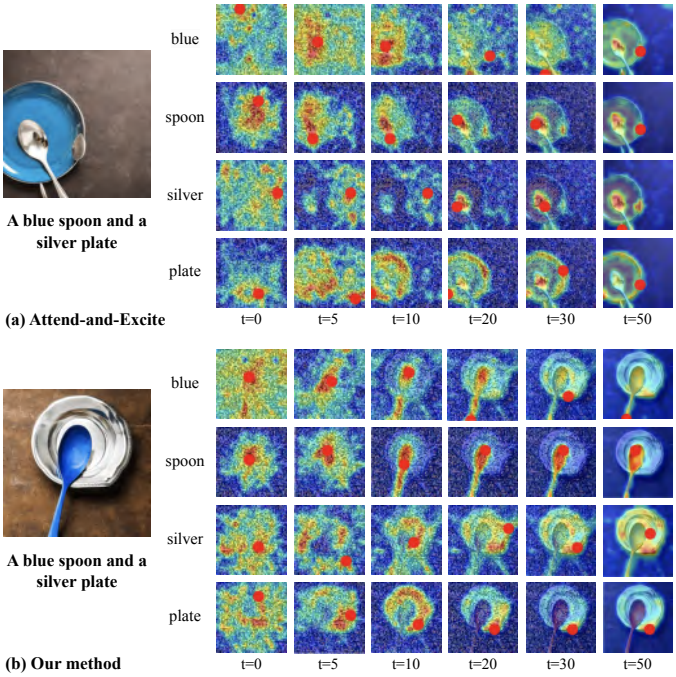

Fig. 2. Cross-attention map visualization for attribute exchange

illustrated in Fig. 2, the Attend-and-Excite method incorrectly assigns attention regions to the "blue" and "silver" tokens, resulting in an inaccurate association between color attributes and objects.

The phenomenon of attention misallocation, which persists in prior research, can be attributed to several key factors. Firstly, the absence of explicit constraints on attention regions and boundaries often results in attention being improperly assigned to incorrect locations. Without explicit guidance, attention may wander to irrelevant areas, potentially leading to suboptimal performance. Secondly, there is a tendency to overlook regions with initially low attention values in the original cross-attention map. This oversight arises from the misconception that low attention values imply insignificance, whereas they may still contain crucial information. Thirdly, during the adjustment of the cross-attention map, the differing importance of nouns and modifiers is not adequately considered. This neglect can lead to imbalanced attention allocation, as nouns and modifiers often carry distinct weights and levels of significance within a given context.

In this study, we propose a novel training-free approach for compositional T2I synthesis using layout-guided diffusion models. Specifically, we employ the Chain-of-Code (CoC) prompting technique of Large Language Models (LLMs) to interpret textual prompts and generate object layouts. Building on these layouts, we introduce two innovative layout-guided loss functions designed to reassign the attention within the cross-attention layers of SD-based models during the denoising process. Firstly, we introduce the Patch-Oriented Cross-Attention (PCA) loss, which emphasizes high activation values for image patches that attend to all tokens in the prompt across the layout. Additionally, as a supplement, the Region-Oriented Cross-Attention (RCA) loss is designed to enhance average attention within the layout, thereby increasing the probability of generating objects within specific regions. Moreover, both the PCA and RCA losses account for the varying significance of nouns and modifiers by distinguishing between their respective loss terms. The incorporation of these losses ensures that our method consistently outperforms multiple strong baselines by a significant margin.

Our main contributions can be summarized as follows:

- We leverage LLMs as layout generators and enhance their performance through CoC prompting. This approach not only improves the accuracy of the generated layouts but also ensures that they more precisely align with the input textual prompts.
- We introduce two novel layout-guided loss functions to reassign cross-attention in SD-based models during sampling. Specifically, our proposed PCA and RCA losses are designed to differentiate the varying significance of nouns and modifiers.
- We conduct a comprehensive experimental evaluation that compares our method with multiple strong baselines, demonstrating its superiority in enhancing the alignment between generated images and a diverse range of textual prompts, while also ensuring high usability as a ready-to-use plugin.

II. Related Work

A. Text-to-Image Diffusion Models

Diffusion models [7]–[10], [15]–[24] have emerged as prominent tools in T2I synthesis due to their remarkable ability to produce high-fidelity images. For instance, GLIDE [17] pioneers T2I methods by incorporating diffusion models and adopting a classifier-free approach. Similarly, Imagen [20] also employs classifier-free guidance for image synthesis. Stable Diffusion (SD) [7] is a leading framework that incorporates diffusion models within latent space. SD uses an autoencoder to transform the images into a lower-dimensional space and diffuses within this latent space, achieving a balance between algorithmic efficiency and image fidelity. DALL·E 2 [19] is another significant work in this domain, comprising a prior for generating CLIP image embeddings based on textual prompts and a decoder for generating images conditioned on these embeddings. Yu et al. [25] introduce ChatGenImage, a versatile tool that leverages LLMs, AIGC models, and label foundation toolkits to generate fine-grained, richly annotated images for data augmentation through model collaboration. Despite the significant advancements made by diffusion models in T2I synthesis, they still face challenges in accurately capturing the semantics of original text, particularly when dealing with multi-object generation or complex scene descriptions.

B. Compositional T2I Synthesis

Recent research has underscored the pivotal role of attention mechanisms in aligning text and images within T2I diffusion models [26]–[32]. Consequently, there's a burgeoning interest in exploring techniques to manipulate attention maps, aiming to enhance the compositional generation capabilities of these models [12]–[14], [32]–[36]. For example, Attend-and-Excite

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

[12] directs a pre-trained diffusion model to optimize cross-attention maps during sampling. While Attend-and-Excite partially mitigates issues like object omission, it still faces challenges such as attribute leakage. On the other hand, Structured Diffusion [14] employs consistency trees to segment prompts into noun phrases, adjusting cross-attention layers in diffusion-based T2I models. However, the results often lack significant modifications that correct the semantic faults. Composable Diffusion [32] creates an image by combining multiple outputs from a pre-trained diffusion model. Each output focuses on capturing distinct image components, which are merged using compositional operators to produce a cohesive image. SynGen [13] uses contrastive loss to enhance the correlation within the same phrase while reducing the attentional correlation between different phrases. Gong et al. [37] propose SimM, a training-free layout calibration system for text-to-image generators. Following a "check-locate-rectify" pipeline, it first determines if rectification is needed by analyzing the prompt and cross-attention maps. If misaligned objects are detected, SimM identifies their activations and relocates them based on dependency parsing and heuristic rules. Shirakawa et al. [38] introduce NoiseCollage, a layout-aware text-to-image diffusion model that generates multi-object images by estimating noises for individual objects separately and merging them during denoising. Wu et al. [39] present the Self-correcting Language-Driven (SLD) framework, which enhances text-to-image alignment through iterative refinement using detectors and LLMs. TokenComposer [40] improves text-to-image consistency by enforcing token-wise alignment between image content and object segmentation maps during finetuning. GLIGEN [16] extends pre-trained text-to-image diffusion models by incorporating grounding inputs. To retain the original model's knowledge, GLIGEN freezes its weights and injects grounding information through new trainable layers using a gated mechanism. In this study, we propose a novel approach to reassign cross-attention in diffusion models by leveraging explicit layout representations.

III. Proposed Method

For SD-based diffusion models, the key to enhancing alignment between text and images lies in the accuracy of attention allocation. Following previous work [12]–[14], we adopt a training-free approach to reassign cross-attention in diffusion models, thereby improving compositional text-to-image synthesis. As shown in Fig. 3, we first employ the CoC prompting technique of LLMs to interpret textual prompts and generate object layouts with spatial coherence. Leveraging these layouts provides explicit constraints and guidance for adjusting attention. Building on these layouts generated by LLMs, we introduce two innovative layout-guided loss functions to refine attention in the cross-attention layers during the denoising process. We begin by presenting background information in Sec. III-A. In Sec. III-B, we illustrate the application of CoC prompting in LLMs for generating layouts. In Sec. III-C, we outline the derivation of the two proposed loss functions.

A. Preliminaries

We apply our method to the publicly available SD model [7], which operates within the latent space of an autoencoder. The encoder $\mathcal{E}(\cdot)$ is trained to transform an image x into a spatial latent code $z = \mathcal{E}(x)$. The decoder $\mathcal{D}(\cdot)$ learns to reconstruct the image from the latent code, aiming for $\mathcal{D}(\mathcal{E}(x)) \approx x$. Then, the conditional diffusion model $\epsilon_\theta(\cdot)$ is trained in the latent space to produce latent codes based on a given textual prompt P. During the training of the diffusion model, a mean-squared reconstruction loss is employed:

$$\mathcal{L}_{sd} := \mathbb{E}_{z\sim\mathcal{E}(x),P,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[||\boldsymbol{\epsilon} - \epsilon_\theta(z_t, t, \tau_\theta(P))||_2^2\right] \quad (1)$$

where $\boldsymbol{\epsilon}$ is drawn from a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, t denotes the time step, z_t represents the latent variable at time t, and $\tau_\theta(\cdot)$ refers to the pre-trained CLIP text encoder [41]. During inference, random Gaussian noise z_T is iteratively denoised to z_0, and the final image is generated by the decoder as $x' = \mathcal{D}(z_0)$.

To integrate text information into the image synthesis process, text embedding features f_p are derived from the pre-trained CLIP text encoder, $f_p = \tau_\theta(P) \in R^{N_k \times N_e}$, where N_k denotes the number of tokens and N_e represents the embedding dimension. The key $\mathbf{K} \in \mathbb{R}^{N_k \times d}$ and value $\mathbf{V} \in \mathbb{R}^{N_k \times d}$ are generated from f_p through projection layers. Given a set of queries $\mathbf{Q} \in \mathbb{R}^{N_q \times d}$ calculated from feature maps of resolution $h \times w$, the cross-attention map is defined as:

$$A = \text{softmax}\left(\frac{\mathbf{QK^T}}{\sqrt{d}}\right) \quad (2)$$

where $N_q = h \times w$ and d denotes the output dimension key and query values. At denoising step t, reshaping and indexing A yields $A_i \in [0,1]^{h' \times w'}$, which represents the attention map between the subject token i and each spatial location in the feature map.

B. Layout Prediction with LLMs

Generating images from text requires a text encoder equipped with robust reasoning capabilities, capable of comprehending the intricate relationships between objects inferred from textual cues. Prior research has predominantly focused on either fine-tuning existing layout generators [31] or training novel text-to-layout generators [28], both of which involve significant overhead in fine-tuning or training. Given the powerful reasoning and analytical capabilities of LLMs, we propose predicting layouts from text using LLMs via the chain of code prompting. The Chain of Code (CoC) [42] is an effective extension of the Chain of Thought (CoT) [43] for enhancing LLMs' code-driven reasoning. Its core concept lies in prompting LLMs to construct semantic sub-tasks within a program as flexible pseudocode. During runtime, these pseudocode segments can then be explicitly identified and processed through a code interpreter and LMulator (LLMs' code emulator). In our implementation, before predicting layouts, we manually provide LLMs with pseudocode and comments relevant to layout generation. This step serves to illustrate the rules and principles underlying layout generation. Subsequently, the LMulator executes the code according to

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

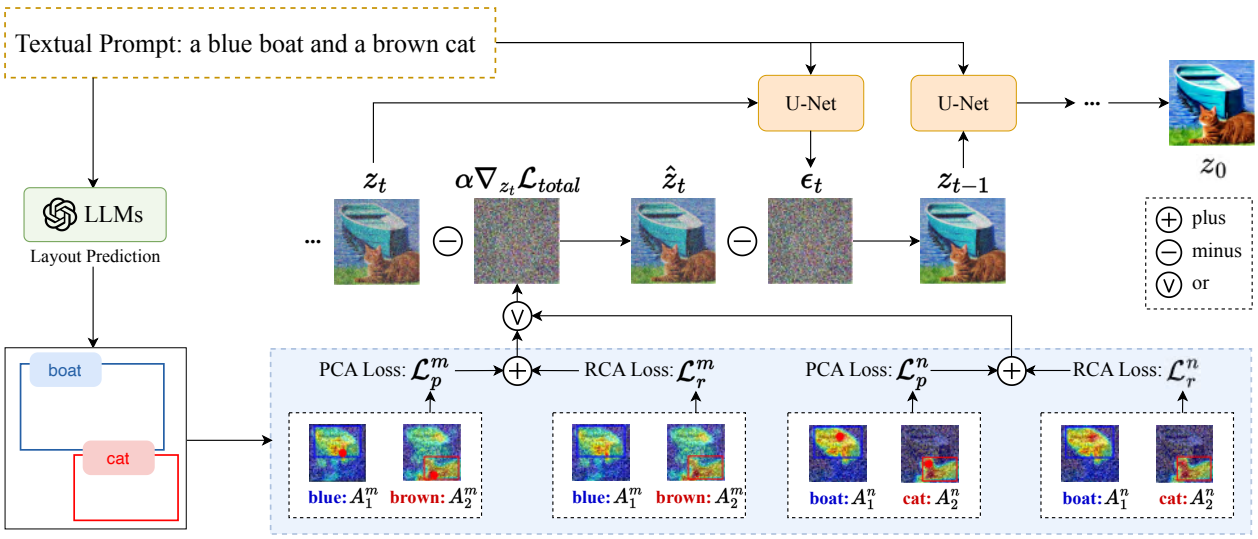

Fig. 3. An overview of our text-to-image generation pipeline. First, we leverage the CoC prompting technique of LLMs to generate object layouts from textual prompts. Next, we utilize the PCA loss and RCA loss to reassign cross-attention in diffusion models during the denoising process.

Prompt: A bed with a quilt, a wardrobe and a desk next to the bed, There is a computer on the desk and two picture on the wall.

```
objects_list = parse_objects(caption)
# objects_list = ["bed", "quilt", "wardrobe", "desk", "computer", "picture"]
for obj in objects_list:  # obj = "quilt" (updated for each loop)
    number = get_number(obj, caption)  # number = 1
    location_description = get_location_description(obj, caption)
    # "The quilt covers the bed, so it should also be to the left of ..."
    objects_attribute_dict[obj] = {"number": number, "location description":
    location_description}
    for id in range(number):  # id = 1
        layout = get_layout(obj, id, location_description)  # layout = (60, 35, 135, 210)
        objects_layout_dict[f"{obj}-{id}"] = layout
```

Fig. 4. An example of CoC prompting for generating layouts.

the guidance provided by the comments, thereby predicting layouts. An example of CoC prompting for generating layouts is depicted in Fig. 4:

Step 1: Parse the objects based on the input textual prompt.

Step 2: Determine the number of objects indicated by the textual prompt.

Step 3: Generate descriptions of each object's location by analyzing their spatial relationships based on the textual prompt.

Step 4: Create layouts for each object by considering the number of objects and their location descriptions. Each layout should be formatted as (x1, y1, x2, y2), where (x1, y1) represents the top-left coordinate and (x2, y2) represents the bottom-right coordinate.

Note that the generated layouts are defined as bounding boxes, with each box corresponding to a textual phrase describing its content. For example, in the phrase "red dog", both "red" and "dog" tokens share the same layout (bounding box). Fig. 5 displays an example of inference results for layout prediction using CoC prompting.

Prompt: A frog on the top of a cup

Attribute Dictionary:

```
{
 "frog": {
   "number": 1,
   "location description": "The location relationship is mentioned; the frog is
on top of the cup. Considering the cup as a base, the frog should be
positioned above it. Since it's on top, it should be located in the upper part of
the picture."
  },
 "cup": {
   "number": 1,
   "location description": "The cup is mentioned to be the base for the
frog. It should occupy a significant portion of the lower part of the picture to
provide a stable base for the frog on top."
  }
}
```

Layout Dictionary:

```
{
 "frog-0": (70, 5, 185, 110),
 "cup-0": (25, 110, 230, 256)
}
```

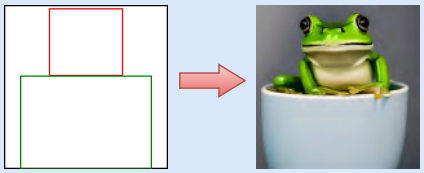

Fig. 5. An example of layout prediction results.

C. Layout-guided Attention Control

In this section, we introduce the Patch-Oriented Cross-Attention (PCA) loss and the Region-Oriented Cross-Attention (RCA) loss, which reassign attention in the cross-attention layers of SD-based models during the denoising process. Specifically, PCA and RCA losses address the varying importance of nouns and modifiers by differentiating the associated loss terms.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

1) Patch-Oriented Cross-Attention Loss: The PCA loss emphasizes high activation values for image patches attending to all tokens in the prompt. This emphasis is crucial because low attention values in these regions can lead to compositional challenges like object omission or attribute leakage. For a specific textual prompt like "two black dogs", both the token "dog" and the token "black" are associated with multiple (two) layouts. The PCA loss first identifies patches with the highest attention values for each token within these layouts. Then, it selects the minimum attention value from these patches with maximum attention. Ultimately, the sum of these minimum attention values across all tokens is computed. Additionally, the PCA loss distinguishes between loss terms for nouns and modifiers during optimization.

The PCA loss tailored for nouns is primarily utilized to ensure the generation of objects, defined as follows:

$$\mathcal{L}_p^{(n)} = 1 - \frac{1}{k_n} \sum_{i \in G_n} \min_{q \in B_i} \{\max[A_i^n \odot M_i^q]\} \tag{3}$$

where G_n denotes a set comprising k_n noun tokens, B_i refers to the associated layouts of token $i \in G_n$, A_i^n represents the cross-attention map of token i at time t, and M_i^q represents the binary mask of the q-th layout in B_i. Here, max calculates the maximal patch attention value within layout q for the cross-attention map A_i^n.

The PCA loss tailored for modifiers is employed to optimize the attributes associated with the objects, defined as follows:

$$\mathcal{L}_p^{(m)} = 1 - \frac{1}{k_m} \sum_{j \in G_m} \min_{q \in B_j} \{\max[A_j^m \odot M_j^q]\} \tag{4}$$

where G_m denotes a set comprising k_m modifier tokens, B_j refers to the associated layouts of token $j \in G_m$, A_j^m represents the cross-attention map of token j at time t and M_j^q represents the binary mask of the q-th layout in B_j. Here, max calculates the maximal patch attention value within layout q for the cross-attention map A_j^m.

2) Region-Oriented Cross-Attention Loss: When certain subject tokens within their associated layout have low initial attention values in the highest attention patches of the original cross-attention map, the PCA loss may not effectively boost attention values, leading to issues such as object omission. To address this, we propose the RCA loss to increase the average attention within the layout. This enhancement aims to improve object generation likelihood within the region during the denoising process and reduce problems related to object omission. Additionally, the RCA loss differentiates between loss terms for nouns and modifiers during optimization.

The RCA loss tailored for nouns is primarily utilized to enhance the generation of objects, defined as follows:

$$\mathcal{L}_r^{(n)} = 1 - \frac{1}{k_n} \sum_{i \in G_n} \min_{q \in B_i} \{mean[A_i^n \odot M_i^q]\} \tag{5}$$

Here, $mean$ calculates the average attention value within the q-th layout for the cross-attention map A_i^n and other symbols remain consistent with Eq. (3).

The RCA loss tailored for modifiers is employed to improve the attributes associated with the objects, defined as follows:

$$\mathcal{L}_r^{(m)} = 1 - \frac{1}{k_m} \sum_{j \in G_m} \min_{q \in B_j} \{mean[A_j^m \odot M_j^q]\} \tag{6}$$

Here, $mean$ calculates the average attention value within the q-th layout for the cross-attention map A_j^m and other symbols remain consistent with Eq. (4).

3) Sampling with Cross-Attention Losses: Based on the PCA and RCA losses, during the optimization process, we first minimize two noun losses to optimize the noise sample z_t at each step, prioritizing object generation in the early denoising stages. When the two noun losses fail to meet the threshold conditions, we then minimize two modifier losses to optimize the noise sample z_t, aiming to refine the attributes associated with the objects. If the modifier losses also fail to meet the threshold conditions, we revert to minimizing the noun losses and alternate between the two types of losses as needed. The update process is illustrated in Fig. 3 with the following formula:

$$\hat{z}_t \leftarrow z_t - \alpha \nabla_{z_t} \mathcal{L}_{total} \tag{7}$$

where α represents the step size that controls the impact of the optimization during denoising. $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \begin{cases} \mathcal{L}_p^{(n)} + \mathcal{L}_r^{(n)}, & (\mathcal{L}_p^{(n)} \geq 1 - \gamma_p^{(n)}) \vee (\mathcal{L}_r^{(n)} \geq 1 - \gamma_r^{(n)}) \\ \mathcal{L}_p^{(m)} + \mathcal{L}_r^{(m)}, (\mathcal{L}_p^{(m)} \geq 1 - \gamma_p^{(m)}) \vee (\mathcal{L}_r^{(m)} \geq 1 - \gamma_r^{(m)}) \end{cases} \tag{8}$$

where $\gamma_p^{(n)}$ denotes the threshold for $\mathcal{L}_p^{(n)}$, $\gamma_r^{(n)}$ denotes the threshold for $\mathcal{L}_r^{(n)}$, $\gamma_p^{(m)}$ denotes the threshold for $\mathcal{L}_p^{(m)}$, and $\gamma_r^{(m)}$ denotes the threshold for $\mathcal{L}_r^{(m)}$. Observations suggest that attention regions for noun tokens are relatively concentrated, while those for modifier tokens are relatively dispersed. Therefore, we assign a higher threshold for noun losses and a lower threshold for modifier losses. To prevent a deterioration in the quality of produced images due to continuous optimization strategies throughout the denoising process, we restrict the update of the noise sample z_t to the first t' steps, where $t' < t$, with t' representing the number of steps using noun and modifier losses.

IV. Experiment

A. Implementation Details

Evaluation Setup. First, we use the GPT-3.5 model through its API to generate bounding boxes from textual prompts. To ensure accurate layout generation using LLMs with CoC prompting, we provide comprehensive guidance and examples, including details such as maximum bounding box size and desired output format, to facilitate optimal layout generation. Next, we apply our method to the publicly available SD model [7] and use the PCA loss and RCA loss on cross-attention maps with a resolution of 16×16. Note that Gaussian smoothing is applied to the cross-attention map before optimization, as described in [12]. The step size is $\alpha = 1$. All images are generated with 50 denoising steps, while the number of steps applying our losses t' is set to 30. Please refer to the supplementary materials for further implementation details.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

Baselines. We perform a thorough comparative analysis of our proposed method against several state-of-the-art approaches for generating images from complex textual descriptions. The baseline methods include: SD [7], Structured Diffusion [14] (referred to as Structured), Composable Diffusion [32] (referred to as Composable), Attend-and-Excite [12] (referred to as Attn-Exct), Layout-Diffusion [33], Divide-Bind [34], Initno [35], Conform [36], SynGen [13], TokenCompose [40] and GLIGEN [16]. In our analysis, both our method and the baselines use the SD models. We utilize SD version 1.4 (SD v1.4) and SD version 2.1 (SD v2.1) as the foundational models for this comparison. Additionally, we compare our method with two state-of-the-art text-to-image models: FLUX [44] and Stable Diffusion 3.5 (SD v3.5) [45]. For evaluation, we use the FP8-quantized FLUX.1 Schnell and the Medium version of SD v3.5.

Benchmark. Previous research has explored specific aspects of compositional text-to-image generation and established benchmarks for evaluation. For instance, datasets such as Attn-Exct [12], CC-500 [14], and ABC-6K [14] are designed primarily for evaluating color attribute binding. On the other hand, although HRSBench [46] offers a comprehensive evaluation benchmark for T2I models, including metrics for accuracy, robustness, generalization, fairness, and bias, it does not fully cover the scope of compositional text-to-image generation. Consequently, after a thorough comparison, we choose T2I-CompBench [47] for evaluation because it is a comprehensive benchmark tailored for compositional text-to-image generation. This benchmark consists of 6,000 compositional textual prompts across diverse categories such as color binding, shape binding, texture binding, spatial relationships, non-spatial relationships, and complex compositions. Attn-Exct [12] has highlighted the interconnected nature of issues like attribute exchange, attribute leakage, object omission, and object leakage. Based on this insight and in accordance with the evaluation metrics of T2I-CompBench, we use the BLIP-VQA metric [48] to thoroughly assess these aspects. In addition, we employ the UniDet-based metric [49] for evaluating spatial relationships and the CLIP score [50], [51] for assessing interaction relationships.

B. Quantitative Analysis

Table I displays the quantitative comparison results between our method and these baseline methods. The analysis of Table I reveals several key insights: (1) The generation capability of different SD versions influences compositional generation performance. Across both baselines and our method, models built on SD v2.1 consistently outperform those based on SD v1.4. (2) Our proposed method outperforms all baseline methods on the BLIP-VQA metric, except for FLUX and SD v3.5. While our method does not achieve the highest CLIP score, we suspect that the CLIP score may not fully capture interaction relationships. Several studies [46], [52], [53] highlight that, while the CLIP score provides a strong average semantic representation, it often fails to accurately assess attribute binding to objects, especially in complex multi-object scenarios. (3) To further demonstrate the generalization capability of our PCA and RCA losses, we assess the effectiveness of our proposed plugin solutions by evaluating the performance of baseline models with and without these losses. Specifically, we integrate our losses into Attn-Exct [12], Conform [36], and SynGen [13]. Our findings show that this integration significantly improves attribute binding in these baseline methods. However, incorporating both losses simultaneously degrades the spatial relationships of the baselines. We infer that the losses from our method and the baselines may conflict when addressing spatial tasks. Additionally, we performed two ablation experiments to evaluate the impact of the proposed PCA and RCA losses. In these two experiments, Attn-Exct + PCA integrates the PCA loss into Attn-Exct's original losses, while Attn-Exct + RCA incorporates the RCA loss. As shown in Table I, both of our proposed losses significantly improve Attn-Exct's performance in terms of attribute binding. These results highlight the superior generalization ability of our method compared to the baselines. (4) On spatial relationships, our method underperforms compared to GLIGEN and TokenCompose. This is primarily because the effectiveness of training-free algorithms in spatial relationships depends heavily on the base model's performance. In contrast, GLIGEN enhances spatial understanding by integrating layout conditions with text phrases through a gated self-attention layer. TokenCompose, on the other hand, fine-tunes Stable Diffusion by jointly optimizing the denoising U-Net of the diffusion model with both its original denoising and the grounding objective. (5) Our method demonstrates competitive performance and outperforms FLUX and SD v3.5 in specific aspects, such as shape binding and texture binding. A key advantage of our approach is its plug-and-play nature. It can be seamlessly integrated into other models without additional training.

Furthermore, we selected three representative baselines (i.e., Attn-Exct, Conform, and SynGen) and evaluated them using MLLMs. Following T2I-CompBench++ [54], we employ MLLMs as an evaluation metric. Specifically, the generated images are fed into the MLLMs (in our case, Qwen2.5-VL), which are prompted with two Chain-of-Thought questions: "describe the image" and "predict the image-text alignment score". The Qwen2.5-VL model then returns a score from 0 to 100 based on predefined criteria. Table II presents the quantitative comparison results based on MLLM evaluations. Although Tables I and II use different evaluation metrics, the results are largely consistent. For example, our method achieves the best performance in attribute binding and complex compositions but underperforms compared to the baselines in spatial and non-spatial relationship understanding. Furthermore, incorporating PCA + RCA loss into the baseline methods significantly improves their attribute binding performance.

C. Qualitative Comparisons

Fig. 6 illustrates the qualitative comparison between our method and these baseline methods. Conform and SynGen achieve competitive results both qualitatively and quantitatively. In contrast, other baseline methods struggle with various issues when handling diverse textual prompts, including object omission, object leakage, attribute exchange, attribute leakage, spatial misalignment, and interaction neglect.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

TABLE I
QUANTITATIVE EVALUATION RESULTS, INCLUDING THIRTEEN BASELINES AND FIVE ABLATED VARIANTS OF OUR METHOD.

Method	Color Binding	Shape Binding	Texture Binding	Spatial Relationships	Non-spatial Relationships	Complex
	BLIP-VQA ↑	BLIP-VQA ↑	BLIP-VQA ↑	Unidet ↑	CLIP ↑	BLIP-VQA ↑
SD v1.4 [7]	0.3668	0.3633	0.4074	0.1134	0.3101	0.3553
Structured v1.4 [14]	0.3732	0.3602	0.4020	0.1117	0.3085	0.3404
Composable v1.4 [32]	0.3743	0.3642	0.4058	0.1151	0.3082	0.3514
Attn-Exct v1.4 [12]	0.4732	0.3913	0.5646	0.1133	0.3096	0.3574
Layout-Diffusion v1.4 [33]	0.3489	0.3605	0.3936	0.1086	0.3032	0.3796
Divide-Bind v1.4 [34]	0.4874	0.4070	0.4774	0.1526	0.3087	0.3993
TokenCompose v1.4	0.4557	0.4029	0.5197	0.1855	0.3132	0.3695
Conform v1.4 [36]	0.6844	0.4886	0.6772	0.1388	0.3073	0.4288
SynGen v1.4 [13]	0.6994	0.4736	0.6746	0.1238	0.3084	0.4131
GLIGEN v1.4 [16]	0.3902	0.3717	0.4534	0.1806	0.3012	0.3307
Our Method v1.4	0.6022	0.4563	0.6487	0.1129	0.3060	0.4050
SD v2.1 [7]	0.5450	0.4348	0.5103	0.1565	0.3116	0.3576
Composable v2.1 [32]	0.4837	0.4057	0.4732	0.1553	0.3093	0.3934
Attn-Exct v2.1 [12]	0.5311	0.4288	0.5443	0.1617	0.3099	0.3548
Layout-Diffusion v2.1 [33]	0.4526	0.4138	0.4511	0.1580	0.3034	0.3797
Divide-Bind v2.1 [34]	0.5073	0.4107	0.5448	0.1526	0.3067	0.3950
TokenCompose v2.1 [40]	0.5650	0.4802	0.6033	0.1727	**0.3177**	0.4049
Initno v2.1 [35]	0.4747	0.3975	0.4630	0.1428	0.3050	0.3866
Conform v2.1 [36]	0.7494	0.5164	0.6939	0.1852	0.3095	0.4539
SynGen v2.1 [13]	0.7438	0.5358	0.6855	0.1656	0.3095	0.4384
Our Method v2.1	0.6451	0.4999	0.6703	0.1631	0.3087	0.4319
Attn-Exct v2.1 + PCA + RCA	0.6320	0.5074	0.6558	0.1527	0.3065	0.4169
Conform v2.1 + PCA + RCA	0.7630	0.5368	0.6480	0.1514	0.3086	**0.4587**
SynGen v2.1 + PCA + RCA	0.7743	**0.5789**	**0.7212**	0.1527	0.3074	0.4379
Attn-Exct v2.1 + PCA	0.6014	0.4931	0.6347	0.1505	0.3060	0.4258
Attn-Exct v2.1 + RCA	0.6172	0.4920	0.6527	0.1505	0.3072	0.4176
SD v3.5	**0.7936**	0.5759	0.7178	**0.2869**	0.3128	0.4408
FLUX	0.7510	0.5733	0.6871	0.2608	0.3130	0.4410

TABLE II
QUANTITATIVE COMPARISON RESULTS BASED ON MLLM EVALUATIONS. A HIGHER SCORE INDICATES BETTER PERFORMANCE.

Method	Color Binding	Shape Binding	Texture Binding	Spatial Relationships	Non-spatial Relationships	Complex
Attn-Exct v2.1 [12]	36.81	23.70	33.96	71.24	**79.43**	66.03
Conform v2.1 [36]	58.79	26.07	49.75	**77.24**	78.73	69.34
SynGen v2.1 [13]	58.86	27.78	48.67	76.12	78.90	68.17
Our Method v2.1	51.94	24.53	49.46	66.78	77.25	68.17
Attn-Exct v2.1 + PCA + RCA	48.55	23.63	46.33	66.64	76.77	68.98
Conform v2.1 + PCA + RCA	58.79	26.66	44.62	66.75	79.00	**70.41**
SynGen v2.1 + PCA + RCA	**61.39**	**29.87**	**54.56**	66.66	77.66	67.30

Moreover, Fig. 7 provides a visual comparison between layouts directly generated by LLMs and those generated with CoC prompting. The left side of the figure displays layouts and corresponding images generated without CoC prompting, while the right side shows those generated with CoC prompting. We observe that layouts generated without CoC prompting may have inaccurate spatial relationships, whereas CoC prompting ensures accuracy in these relationships. Furthermore, layouts generated without CoC prompting might omit certain objects. For instance, in the second row, the layout generated without CoC prompting fails to include the gown, whereas CoC prompting ensures that all objects are included.

D. Ablation Study

We conducted several ablation experiments to assess the effectiveness of the proposed PCA loss and RCA loss: (1) w/o PCA Loss: a model variant that excludes the PCA loss from the optimization process. (2) w/o RCA Loss: a model variant that excludes the RCA loss from the optimization process. (3) w/o Distingd Loss: a model variant that does not differentiate between the loss terms for nouns and modifiers in the PCA loss and RCA loss during optimization. (4) + Self-attention: a model variant that considers both cross-attention and self-attention losses during optimization. Specifically, self-attention losses are incorporated similarly to cross-attention losses, encompassing both the PCA loss and RCA loss. For additional details, please refer to the supplementary materials. (5) w/o CoC Prompting: a model variant without CoC Prompting, used to assess the effectiveness of CoC Prompting during layout generation.

The quantitative evaluation results are presented in Table III. Our analysis demonstrates that our model consistently outperforms all other variants across all metrics. From these findings, we can draw several conclusions: (1) Both the PCA loss and RCA loss contribute to performance enhancement to varying extents. (2) The differentiation between loss terms for nouns and modifiers within PCA and RCA losses also contributes

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

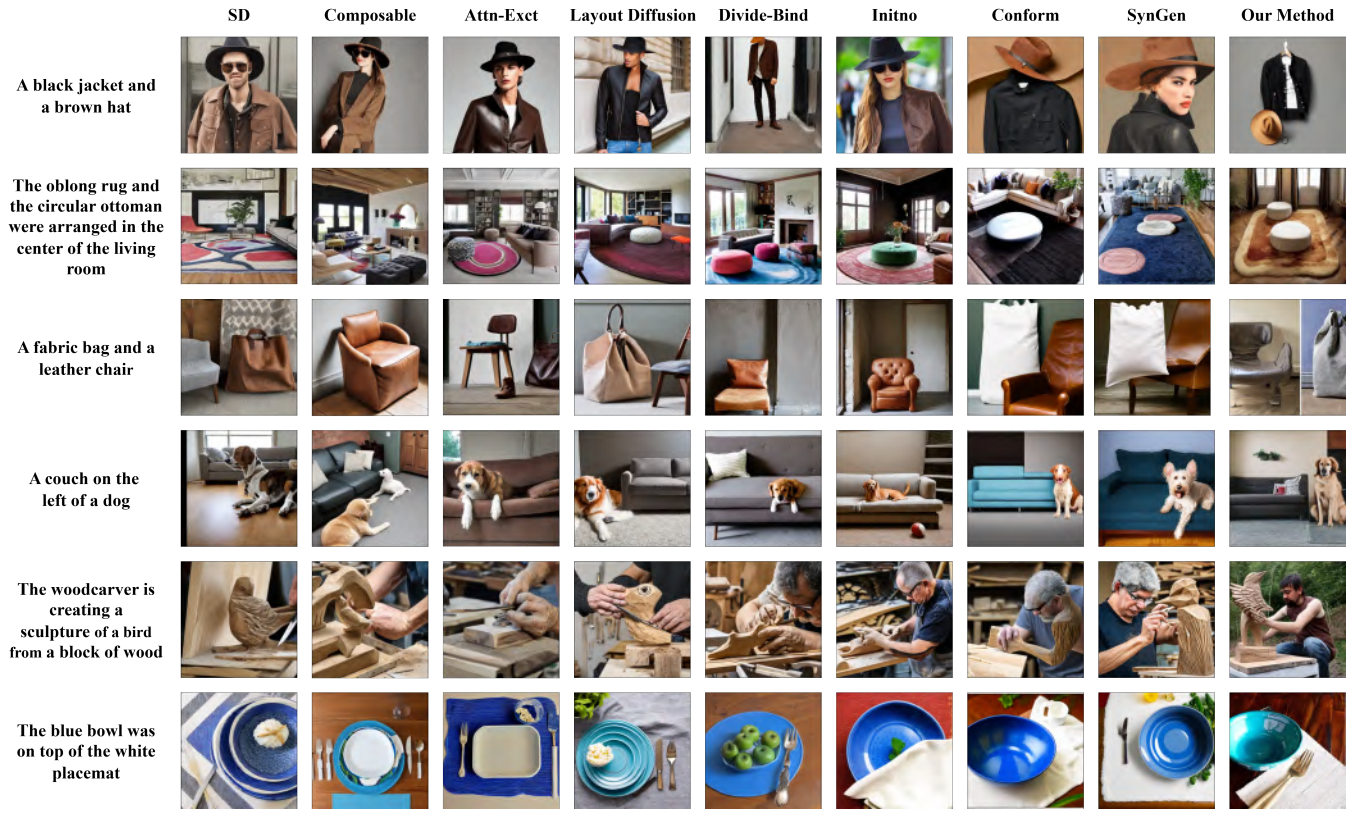

Fig. 6. Qualitative comparison between the baseline methods and the proposed method.

TABLE III
THE QUANTITATIVE ABLATION RESULTS. OUR METHOD CONSISTENTLY OUTPERFORMS ALL OTHER VARIANTS ACROSS ALL METRICS.

Method	Color Binding	Shape Binding	Texture Binding	Spatial Relationships	Non-spatial Relationships	Complex
	BLIP-VQA ↑	BLIP-VQA ↑	BLIP-VQA ↑	Unidet ↑	CLIP ↑	BLIP-VQA ↑
w/o PCA Loss	0.6256	0.4609	0.6439	0.1496	0.3083	0.4147
w/o RCA Loss	0.6035	0.4922	0.6513	0.1497	0.3084	0.4299
w/o Distingd Loss	0.6256	0.4914	0.6461	0.1463	0.3074	0.4226
+ Self-attention	0.6397	0.4908	0.6686	0.1496	0.3074	0.4199
w/o CoC Prompting	**0.6533**	0.4977	**0.6749**	0.1504	0.3067	0.4238
Our Method v2.1	0.6451	**0.4999**	0.6703	**0.1631**	**0.3087**	**0.4319**

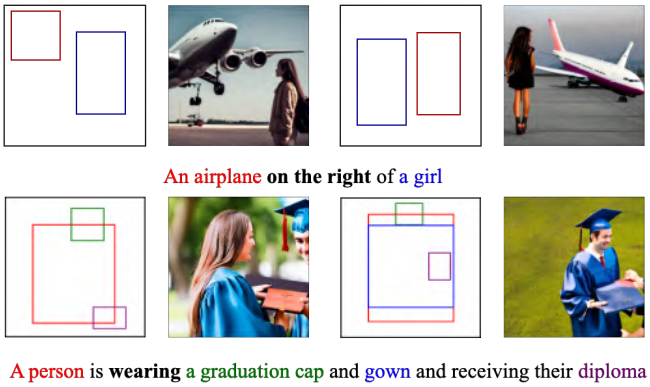

Fig. 7. The visualization of generated layouts.

to performance enhancement. (3) Incorporating both cross-attention and self-attention losses during optimization does not yield performance improvements. (4) While CoC prompting does not substantially improve attribute binding performance, it does enhance spatial capabilities.

The qualitative evaluation results, depicted in Fig. 8, highlight the effectiveness of our method in addressing attribute binding, maintaining object integrity, preserving spatial arrangements, and capturing interaction dynamics, surpassing all other variants. Notably, our method accurately produces items such as the triangular table runner and plastic container, whereas other variants either failed to produce or entirely omitted these objects.

E. User Study

As previously discussed, the CLIP score is not an adequate metric for complex multi-object scenarios, particularly when evaluating interaction relationships. To address this, we conducted two user studies involving 50 participants to assess the effectiveness of our approach in capturing non-spatial relationships.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

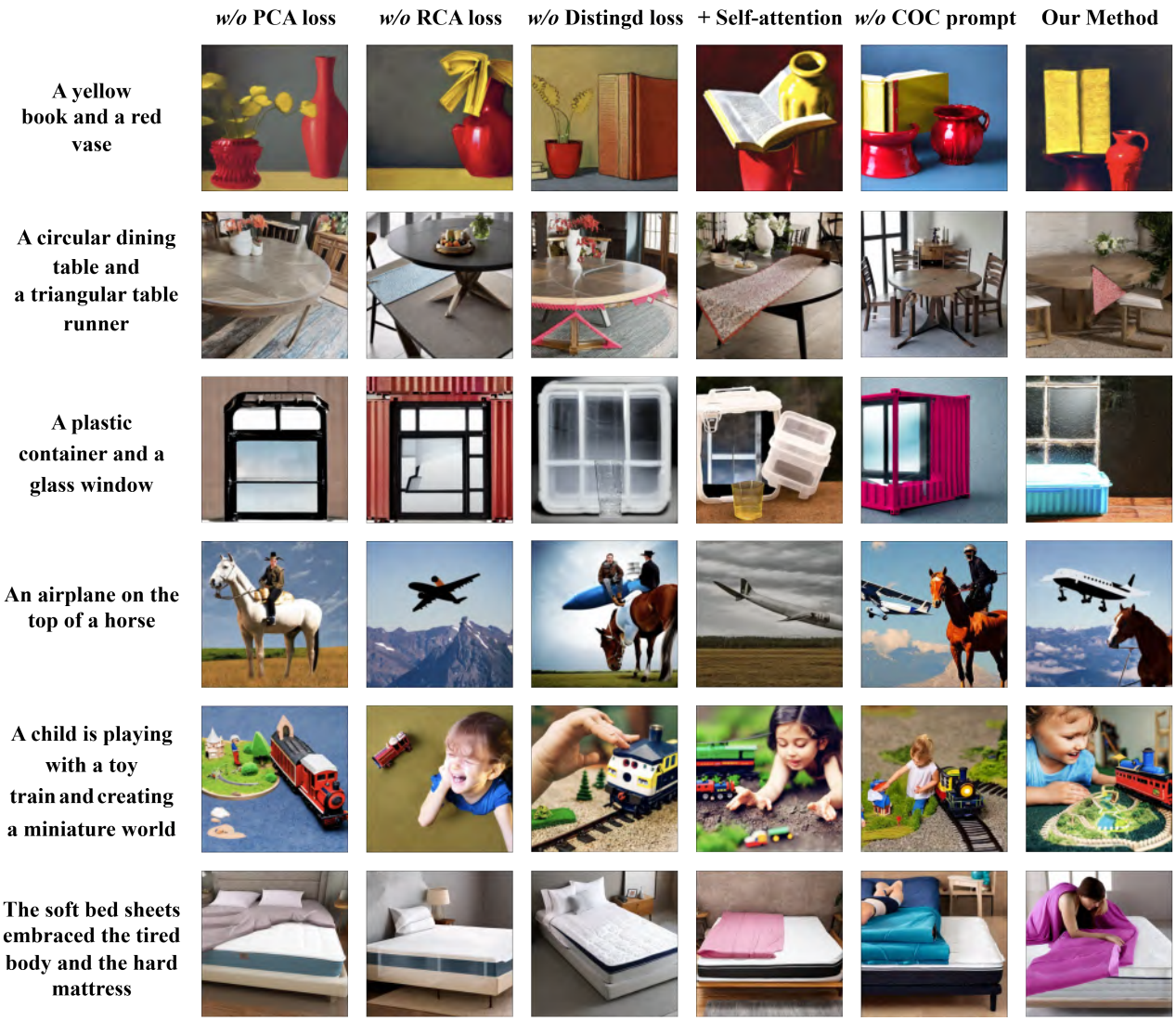

Fig. 8. Qualitative comparison of the ablation results.

In the first user study, we compared our method with ten competitive baselines. We randomly selected 25 textual prompts from the non-spatial relationships test set. For each prompt, two images were generated, one using our method and the other using a baseline method. Each participant was shown the prompt alongside the two images and asked to choose the one that best matched the description. Table IV presents the percentage of participants who preferred our method. The results clearly show that our approach consistently outperforms the baselines.

In the second study, we conducted a rating-based evaluation involving three representative baselines (i.e., Attn-Exct, Conform, and SynGen). Participants rated the generated images based on two criteria: text-image alignment and image fidelity, using a 1-to-5 scale. The prompts and images used were consistent with those in the first study. As shown in Table V, our method achieves higher scores in both evaluation dimensions, further validating its effectiveness in capturing interaction relationships.

TABLE IV
PAIRWISE EVALUATION RESULTS.

Method	**Non-spatial Relationships**
SD v2.1 [7]	73.9%
Composable v2.1 [32]	78.0%
Attn-Exct v2.1 [12]	79.9%
Divide-Bind v2.1 [34]	76.0%
Conform v2.1 [36]	74.9%
SynGen v2.1 [13]	76.6%
GLIGEN v1.4 [16]	81.6%
TokenCompose v2.1 [40]	60.2%
Initno v2.1 [35]	80.8%
Layout-Diffusion v2.1 [33]	75.2%

F. Cross-Attention Map Visualization

In Fig. 9, we present an example of object omission by comparing the cross-attention maps of Attn-Exct v2.1 with those of our method. Given the prompt "A small bathroom with a small brown toilet next to a white sink", Attn-Exct v2.1 initially assigns scattered attention to the brown toilet.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

TABLE V
RATING-BASED EVALUATION RESULTS.

Method	Text-Image Alignment	Image Fidelity
Attn-Exct v2.1 [12]	3.370	2.950
Conform v2.1 [36]	3.652	2.818
SynGen v2.1 [13]	3.414	3.136
Our Method v2.1	**3.804**	**3.48**

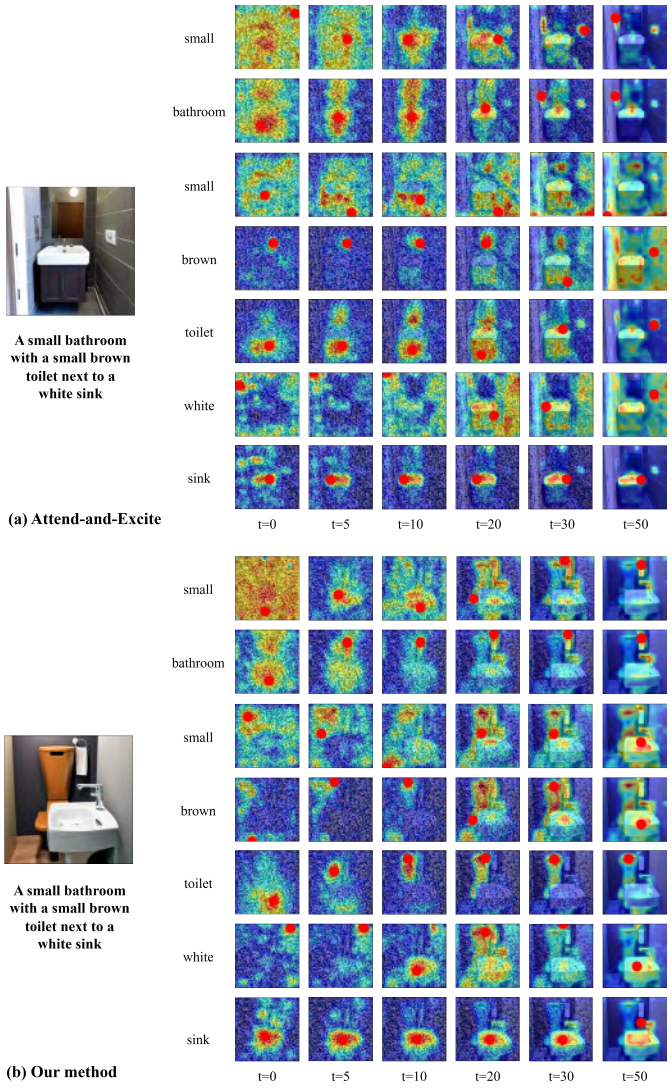

Fig. 9. Cross-attention map visualization for object omission.

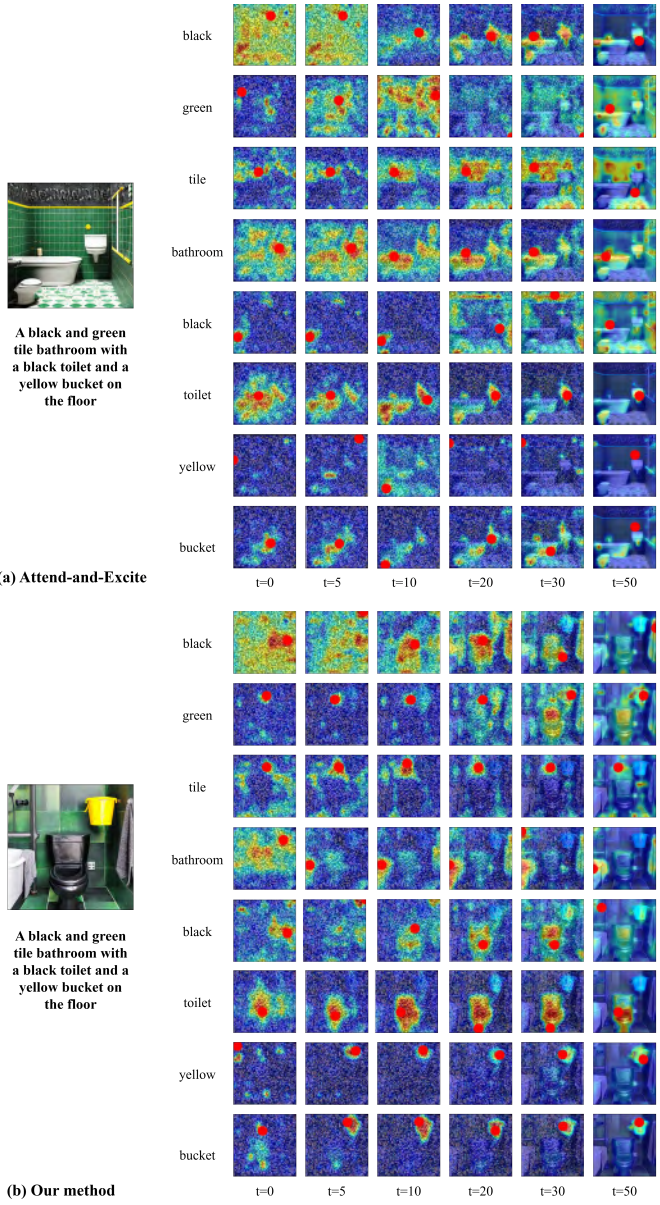

Fig. 10. Cross-attention map visualization for object leakage.

However, during the denoising process, these attention regions are mistakenly redirected toward the sink area, leading to the toilet being omitted in the final output. In contrast, our method also starts with scattered attention but employs PCA and RCA losses to refine the focus, accurately concentrating attention on the toilet and ensuring its correct generation.

In Fig. 10, we present an example of object leakage. Given the prompt "A black and green tile bathroom with a black toilet and yellow bucket on the floor", Attn-Exct v2.1 exhibits overlapping attention regions for the toilet and sink, leading to object confusion. In contrast, our method aligns the attention regions of the black and toilet tokens, as well as the yellow and bucket tokens, effectively preventing object leakage.

In Fig. 11, we present an example of attribute leakage. Given the prompt "A woman in a white shirt and jeans holds a pink umbrella in the rain", Attn-Exct v2.1 misdirects the attention intended for the token pink toward both the shirt and umbrella regions. In contrast, our method correctly focuses attention on the umbrella for pink, thereby preserving the intended attribute information.

G. Limitations

The limitations of our proposed method are illustrated in Fig. 12: (1) *Soft spatial constraint*. In our method, layout guidance serves only as a rough indication of positioning. Thus, the generated images may not entirely conform to the provided layouts. In addition, the initial random noise

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

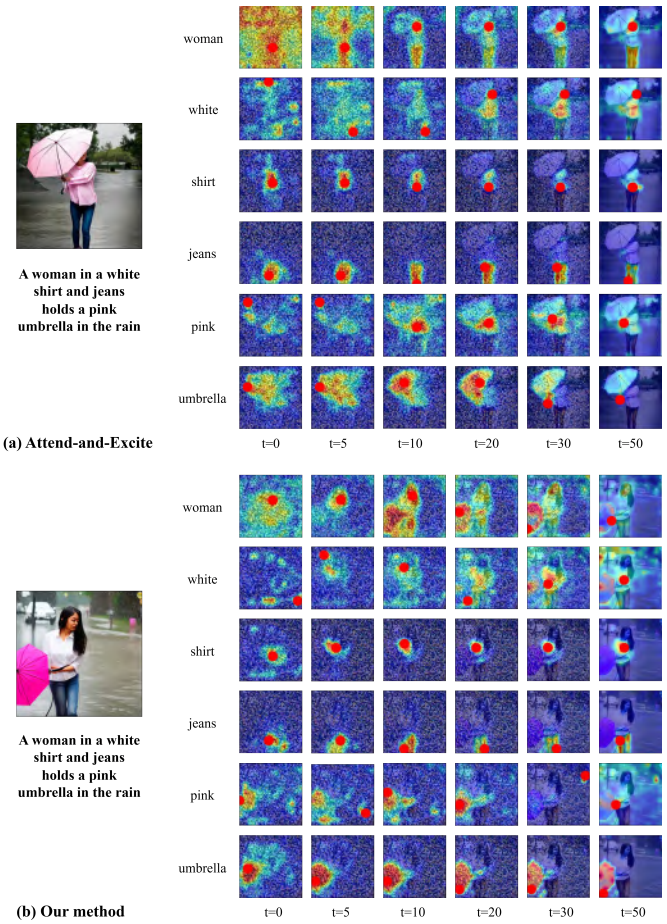

Fig. 11. Cross-attention map visualization for attribute leakage.

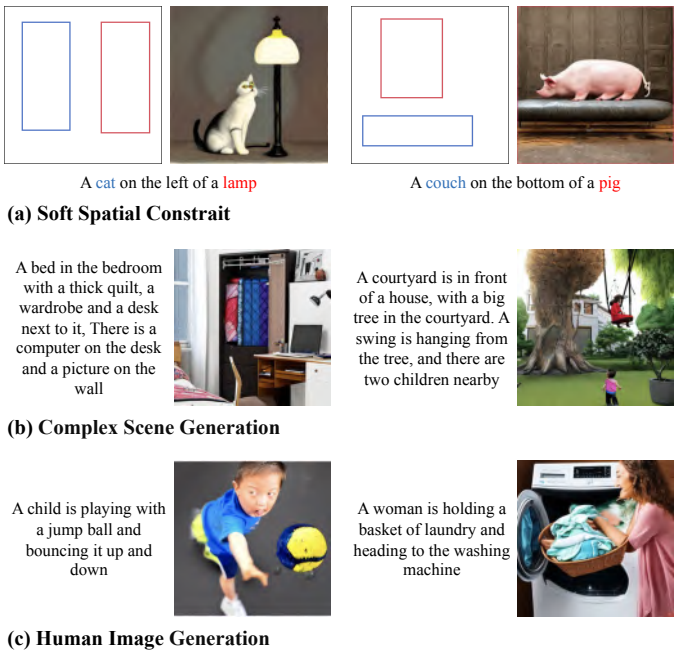

Fig. 12. The limitations of our proposed method.

may sometimes fail to effectively guide the information of a particular token to the desired layout, leading to significant discrepancies between the positioning of the generated image and the layout. (2) *Complex scene generation*. Our method performs well with straightforward textual prompts but struggles when optimizing for complex scenes involving multiple objects and relationships. For instance, when presented with a prompt like "A bed in the bedroom with a thick quilt, a wardrobe, and a desk with a computer on it, as well as a picture on the wall", our method produces unsatisfactory results due to the original SD models' limitations in generating complex scenes. (3) *Human image generation*. When textual prompts involve dynamic actions or the generation of human faces and bodies, poses and facial features often appear distorted. These issues primarily stem from the limited ability of the original SD models to accurately depict humans.

V. CONCLUSION

In this work, we propose a novel, training-free approach for compositional text-to-image synthesis using layout-guided diffusion models. First, we leverage the chain-of-code prompting technique of large language models for generating object layouts directly from textual prompts. We then introduce two innovative layout-guided loss functions: Patch-oriented Cross-Attention (PCA) loss and Region-oriented Cross-Attention (RCA) loss. These loss functions are designed to refine the attention within the cross-attention layers of SD-based models during the sampling process. Extensive experimental evaluations show that our method consistently outperforms multiple strong baselines by a significant margin and ensures high usability as a ready-to-use plugin.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 62471168, 62422204 and 61802100, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LDT23F02025F02 and LY21F020019.

REFERENCES

[1] H. Tan, B. Yin, K. Wei, X. Liu, and X. Li, "Alr-gan: Adaptive layout refinement for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 25, pp. 8620–8631, 2023.
[2] B. Yang, X. Xiang, W. Kong, J. Zhang, and Y. Peng, "Dmf-gan: Deep multimodal fusion generative adversarial networks for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 6956–6967, 2024.
[3] S. Ye, H. Wang, M. Tan, and F. Liu, "Recurrent affine transformation for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 462–473, 2024.
[4] B. Yuan, Y. Sheng, B.-K. Bao, Y.-P. P. Chen, and C. Xu, "Semantic distance adversarial learning for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 1255–1266, 2024.
[5] D. Liu, L. Y. Wu, B. Li, Y. Zhao, Z. Ge, and J. Zhang, "T-person-gan: Text-to-person image generation with identity-consistency and manifold mix-up," *Expert Syst. Appl.*, vol. 288, p. 128178, 2025.
[6] X. Gu, J. Yu, Y. Wong, and M. S. Kankanhalli, "Toward multimodal conditioned fashion image translation," *IEEE Transactions on Multimedia*, vol. 23, pp. 2361–2371, 2021.
[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.
[8] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021, pp. 8162–8171.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3668676

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[10] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021, pp. 8780–8794.

[11] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *CVPR*, 2023, pp. 1921–1930.

[12] 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.

[13] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," in *NeurIPS*, 2023.

[14] W. Feng, X. He, T. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," in *ICLR*, 2023.

[15] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! image synthesis with semantic diffusion guidance," in *WAVC*, 2023, pp. 289–299.

[16] 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 *CVPR*, 2023, pp. 22 511–22 521.

[17] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, vol. 162, 2022, pp. 16 784–16 804.

[18] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[19] 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.

[20] 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," in *NeurIPS*, vol. 35, 2022, pp. 36 479–36 494.

[21] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual representation for controllable image generation," in *CVPR*, 2023, pp. 18 370–18 380.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.

[23] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023, pp. 22 500–22 510.

[24] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro *et al.*, "ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[25] Q. Yu, J. Li, W. Ye, S. Tang, and Y. Zhuang, "Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration," *CoRR*, vol. abs/2305.12799, 2023.

[26] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *CVPR*, 2023, pp. 843–852.

[27] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," in *ICCV*, 2023, pp. 7701–7711.

[28] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," in *AAAI*, 2024, pp. 5544–5552.

[29] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *ICCV*, 2023, pp. 7545–7556.

[30] O. Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor, and D. Cohen-Or, "Localizing object-level shape variations with text-to-image diffusion models," in *ICCV*, 2023, pp. 22 994–23 004.

[31] Q. Wu, Y. Liu, H. Zhao, T. Bui, Z. Lin, Y. Zhang, and S. Chang, "Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis," in *ICCV*, 2023, pp. 7766–7776.

[32] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *ECCV*. Springer, 2022, pp. 423–439.

[33] 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," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[34] Y. Li, M. Keuper, D. Zhang, and A. Khoreva, "Divide & bind your attention for improved generative semantic nursing," in *BMVC*, 2023.

[35] X. Guo, J. Liu, M. Cui, J. Li, H. Yang, and D. Huang, "Initno: Boosting text-to-image diffusion models via initial noise optimization," in *CVPR*, 2024, pp. 9380–9389.

[36] T. H. S. Meral, E. Simsar, F. Tombari, and P. Yanardag, "Conform: Contrast is all you need for high-fidelity text-to-image diffusion models," in *CVPR*, 2024, pp. 9005–9014.

[37] B. Gong, S. Huang, Y. Feng, S. Zhang, Y. Li, and Y. Liu, "Check locate rectify: A training-free layout calibration system for text-to-image generation," in *CVPR*, 2024, pp. 6624–6634.

[38] T. Shirakawa and S. Uchida, "Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging," in *CVPR*, 2024, pp. 8921–8930.

[39] T.-H. Wu, L. Lian, J. E. Gonzalez, B. Li, and T. Darrell, "Self-correcting llm-controlled diffusion models," in *CVPR*, 2024, pp. 6327–6336.

[40] Z. Wang, Z. Sha, Z. Ding, Y. Wang, and Z. Tu, "Tokencompose: Text-to-image diffusion with token-level supervision," in *CVPR*, June 2024, pp. 8553–8564.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139, 2021, pp. 8748–8763.

[42] C. Li, J. Liang, A. Zeng, X. Chen, K. Hausman, D. Sadigh, S. Levine, L. Fei-Fei, F. Xia, and B. Ichter, "Chain of code: Reasoning with a language model-augmented code emulator," in *ICML*, 2024.

[43] X. Chen, Y. Liu, Y. Yang, J. Yuan, Q. You, L.-P. Liu, and H. Yang, "Reason out your layout: Evoking the layout master from large language models for text-to-image synthesis," *arXiv preprint arXiv:2311.17126*, 2023.

[44] B. F. Labs, "FLUX," https://github.com/black-forest-labs/flux, 2024.

[45] Stability-AI, "sd3.5," https://github.com/Stability-AI/sd3.5, 2024.

[46] 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 *ICCV*, 2023, pp. 20 041–20 053.

[47] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," in *NeurIPS*, vol. 36, 2024.

[48] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[49] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *CVPR*, 2022, pp. 7571–7580.

[50] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *EMNLP*, 2021, pp. 7514–7528.

[51] 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 *ICML*, 2021, pp. 8748–8763.

[52] J. Singh and L. Zheng, "Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative VQA feedback," in *NeurIPS*, 2023.

[53] L. W. X. W. Junhui Yin, Xinyu Zhang, "Context-aware prompt learning for test-time vision recognition with frozen vision-language model," *Pattern Recognition*, vol. 162, p. 111359, 2025.

[54] K. Huang, C. Duan, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3563–3579, 2025.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.