Engineering Applications of Artificial Intelligence 153 (2025) 110741

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

MirrorDiff: Prompt redescription for zero-shot grounded text-to-image generation with attention modulation

Chang Liu, Mingwen Shao *, Zhengyi Gong, Xiang Lv, Lingzhuang Meng

State Key Laboratory of Chemical Safety, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

ARTICLE INFO

Keywords:
Text-to-image generation
Large language model
Diffusion model
Attention modulation

ABSTRACT

Large-scale layout-conditioned text-to-image diffusion models have made significant progress and achieved remarkable results in generating diverse and high-quality images, realizing objects appearing in specific regions simultaneously. However, existing methods still fail with attribute coupling, unreasonable spatial relationships expressions and missing objects when the prompt is complex with multiple objects containing multiple attributes. In addition, it is difficult for users to give precise layout conditions for complex prompts. To address the above issues, we propose MirrorDiff, a novel training-free grounded text-to-image-to-text framework by redescription to correct inaccurate content expressions of synthetic images iteratively. Specifically, we first utilize large language models as layout generator which have the ability to understand visual concepts and support plausible arrangements to generate scene layout for complex prompts to help users obtain precision layout more conveniently. Subsequently, to solve small object missing, we design a layout-guided attention modulation strategy to properly adjust attention maps during diffusion generation process, which effectively increases attention of small objects. Additionally, semantic text regeneration supervision is proposed to constrain the redescription to keep consistent with the given text semantically, which aims to mitigate attribute coupling and failures of spatial relationships expressions. We conduct extensive experiments on four benchmarks and our method achieves the best results in all categories on the Holistic, Reliable and Scalable benchmark, which shows that our proposed MirrorDiff achieves state-of-the-art results both quantitatively and qualitatively compared with current superior models.

1. Introduction

In recent years, large-scale diffusion models (Saharia et al., 2022; Rombach et al., 2022; Betker et al., 2023) are used as the primary text-to-image tools to synthesize diverse and semantic images outperforming than Generative Adversarial Networks (Zhang et al., 2021; Tao et al., 2022; Kang et al., 2023; Yang et al., 2024) and autoregression models (Ding et al., 2021, 2022; Ramesh et al., 2022; Chang et al., 2023). However, these methods need large datasets with millions of images and powerful computing resource which are expensive to realize. To mitigate this issue, several fine-tuning (Mou et al., 2024; Zhang et al., 2023) and training-free (Feng et al., 2022; Kim et al., 2023) methods have been proposed. Although the current models have been able to achieve good results, they are unable to precisely generate images that align with text prompt and fail with attribute coupling, unreasonable spatial relationships, missing objects and combinations of the aforementioned problems as shown in Fig. 1, which is primarily due to the low attention response intensity between the main objects and their corresponding tokens. Our work aims to alleviate the above problems by redescription to correct error expressions of synthetic images for zero-shot grounded text-to-image generation.

Recently, Large language models (LLMs) are introduced as layout generators for text-to-image generation (Phung et al., 2024; Feng et al., 2024; Jia and Tan, 2024; Lian et al., 2023; Chen et al., 2023) because they are able to automatically provide scene layouts due to its powerful visual concepts understand and layout arrangement ability. However, LLMs can only provide layout information for each object and are unable to offer high-level spatial relationships that effectively combine the layout information of multiple objects. As a result, the generated images appear to be merely a collage of these objects, failing to express spatial semantic information and attribute content. We discover that it is caused by the inconsistency between the generated image and the original text in the high-level semantic space, which is a representation

* Corresponding author.
E-mail addresses: upc_liuchang@163.com (C. Liu), smw278@126.com (M. Shao), 17865420733@163.com (Z. Gong), lvxiang1997@126.com (X. Lv), lzhmeng1688@163.com (L. Meng).

https://doi.org/10.1016/j.engappai.2025.110741
Received 8 November 2024; Received in revised form 21 February 2025; Accepted 29 March 2025
Available online 22 April 2025

Fig. 1. Large-scale text-to-image diffusion models like Stable Diffusion XL (SDXL) fail to align with given text prompt from different perspectives such as attribute binding, spatial relationship, numeracy, small object generation and complex composition. Our proposed MirrorDiff improves the controllability and accuracy of text-to-image generation.

space focusing on the overall semantics and context where the text is encoded by the pre-trained model. In a word, current training-free generative models combined with LLMs (Phung et al., 2024; Jia and Tan, 2024; Lian et al., 2023) fail to allow for controllable and accurate generation.

To address the issues, we propose MirrorDiff, a novel training-free grounded text-to-image-text framework by prompt redescription which is different from previous structures shown in Fig. 2. The pipeline of MirrorDiff consists of three stages. Specifically, in the first stage, we utilize GPT-4.0 (Achiam et al., 2023) as a layout generator to help users provide a reasonable layout according to given text prompt. In the second stage, to mitigate the issue of disappearance of small objects, we design a layout-guided attention modulation method to adaptively modulate attention map by considering the areas of bounding box of main objects. In the third stage, we design a semantic text regeneration supervision to constrain regenerated text and the original text aligning semantically by reducing the discrepancy between the original text and the regenerated text in high-level semantic space. This aims to alleviate the phenomenon of high-level semantic information missing and splicing objects easily according to layout. The quantitative and qualitative experiments on four benchmarks suggest that our model achieves state-of-the-art results.

The main contributions of this work are as follows:

- we propose MirrorDiff, a novel zero-shot grounded text-to-image-text framework with Large language models by prompt redescription to improve the controllability and accuracy of generative baselines.
- We design a layout-guided attention modulation by adjusting attention maps to mitigate the disappearance of small object effectively.
- A semantic text regeneration supervision is presented to constrain regenerated text and the original text align in high-level semantic space to solve attribute coupling and failures of spatial relationships expressions.
- Extensive experiments on four benchmark datasets demonstrate the effectiveness of our proposed MirrorDiff framework for compositional text-to-image generation.

2. Related works

2.1. Text-to-image diffusion models

The development of high-quality and semantically consistent text-to-image synthesis (Yu et al., 2022; Podell et al., 2023; Saharia et al., 2022; Wang et al., 2023) has been notably advanced by significant breakthroughs in large-scale diffusion models, particularly with the advent of Stable Diffusion (Rombach et al., 2022), while they are supported by substantial data and training resources. However, the variances among data from different domains have weakened the performance of foundational models. Subsequently, several training-free and zero-shot text-to-image diffusion models (Xie et al., 2023; Couairon et al., 2023; Chen et al., 2024) have been proposed which update the latent by introducing some constraints to align attention map with spatial conditions such as image layout and segmentation. Despite encouraging outcomes, existing methods struggle to synthesis semantic and high quality images accurately according to complicated text prompt. On the contrary, our proposed MirrorDiff can mitigate the above problems effectively by prompt redescription.

2.2. Layout-conditioned text-to-image generation

In order to generate objects in specific regions, the layout is injected into large-scale text-to-image models as additional conditional information to guide generation process. Subsequently, the fine-tuning approaches are proposed (Avrahami et al., 2023; Zhang et al., 2023; Li et al., 2023; Yang et al., 2023) which require pairs of box-image data for training, which is time-consuming and labor-intensive. For example, GLIGEN (Li et al., 2023) proposes a trainable gated self-attention layer to incorporate spatial information to stable diffusion for open-set grounded text-to-image generation. However, GLIGEN is not able to express the attributes of the objects correctly. On the contrary, benefiting from our proposed semantic text regeneration supervision, MirrorDiff not only correctly expresses the spatial relationship, but also generates the correct attributes. Subsequently, Reco (Yang et al., 2023) proposes an effective regional control method and inserts an extra set of position tokens containing spatial coordinates into text as input conditions. In addition, several training-free methods focus on improving the

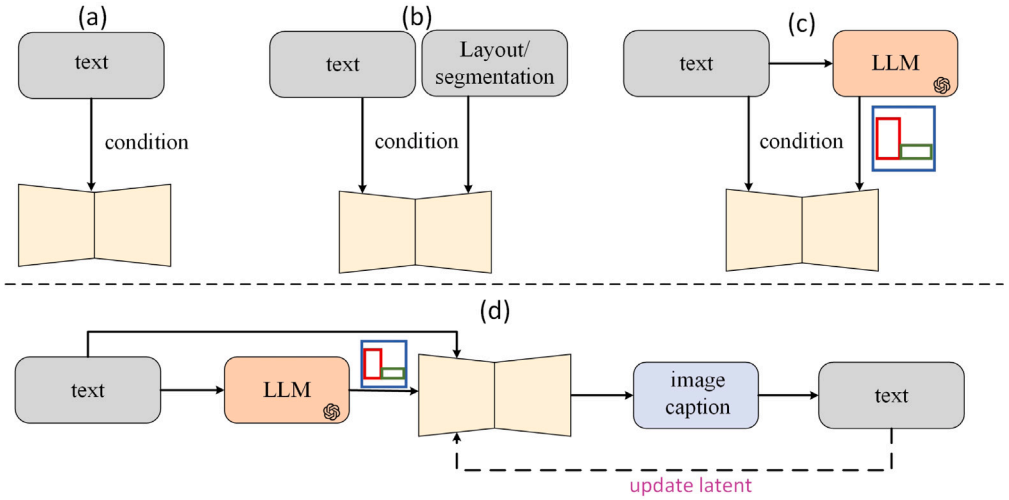

Fig. 2. Architecture comparison between (a) text-guided diffusion models (Podell et al., 2023), (b) layout-guided diffusion models (Li et al., 2023), (c) text/layout-to-image diffusion model with Large language models (Phung et al., 2024), (d) our MirrorDiff.

controllability by introducing additional constraints which requires no extra data and training source. For instance, Boxdiff (Xie et al., 2023) proposes Box-Constrained Diffusion model to ensure objects appearing in the specified location and restricting the scales by introducing inner-box constraint, outer-box constraint and corner constraint. In addition, DenseDiffusion (Kim et al., 2023) first demonstrates the relationship between image layout and attention maps and modulate intermediate attention maps in pre-trained model but neglects the generation of small objects. In summary, despite notable improvements, these methods require complicated layout information which is difficult to obtain for users. While in this paper, we adequately explore the ability of visual concept understand and layout generation of LLM to help users support precision layout.

2.3. LLMs for text-to-image generation

In recent years, LLMs have demonstrated superior performance in cross-modal tasks which is primarily attributed to their robust in-context learning and reasoning ability. Inspired by these developments, several studies have harnessed the power of LLMs to generate image layouts, thereby enhancing the performance of text-to-image models. LayoutGPT (Feng et al., 2024) introduces an innovative approach that generates in-context visual demonstrations through a style sheet language, thereby augmenting the visual planning capabilities of LLMs. But the generated images from LayoutGPT fail to be consistent with the original text in the high-level semantic space, which results in misalignment with text prompt. In contrast, our MirrorDiff designs semantic text regeneration supervision to correct errors in the generated images by minimizing the differences between the generated text and the original text in the semantic space, which enhances semantic consistency between generated images and input text. Attention-refocusing (Phung et al., 2024) leverages GPT (Achiam et al., 2023) to create layouts from text conditions for grounded text-to-image model generation which is similar to LLM-grounded Diffusion (Lian et al., 2023). Subsequently, DivCon (Jia and Tan, 2024) introduces a divide-and-conquer strategy that segregates the processes of reasoning and visual planning to independently enhancing the precision of generated layout. Blob-GEN (Nie et al., 2024) propose a blob-grounded text-to-image diffusion model for compositional generation which leverage the compositionality of LLMs generate blob representations from text prompts rather than layout. However, the results of aforementioned models seems to be the simple concatenation of individual objects neglecting the expression of high-level semantic features and expressing attribute information uncontrollably. While the semantic text regeneration supervision we proposed is capable of correcting error results and refining the generated image effectively.

3. Methodology

3.1. Preliminary

Text-to-Image Diffusion Models. Diffusion model generates an image from a random noise map z_t by constantly subtracting the predicted noise through a trainable time step dependent noise prediction network $\epsilon_\theta(z_t, t)$, where t is the time step. While text-to-image diffusion models like stable diffusion utilize Variational Autoencoder consisting of an encoder $\mathcal{E}$ and a decode $\mathcal{D}$ to continue diffusion process in a low dimensional space. Latent feature z_t encoded from image as query Q interacts with key K from text embedding c, which is vector representation capturing their semantic information in a continuous vector space and is encoded by a pre-trained CLIP encoder ρ. Now the noise predictor $\epsilon_\theta = (z_t, t, \rho(c))$ in U-Net architecture and optimized by the following:

$$\mathcal{L} = \mathbb{E}_{Z\sim\varepsilon(x);\rho(c);\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \rho(c)\right)\right\|_2^2\right]. \tag{1}$$

Attention Layer. There are many attention layers in the text-to-image diffusion model, mainly including cross-attention layer and self-attention layer. Specifically, text prompt is encoded into text embedding with a CLIP (Radford et al., 2021) text encoder and obtain key and value via linear mapping networks, then interacts with a set of queries computed from image features to achieve an attention map which defined as below:

$$A = softmax(\frac{QK^T}{\sqrt{d}}), \tag{2}$$

where Q is query from intermediate image features and K is key from text embedding, d denotes the dimension of query and key which is used to normalize the softmax values.

While in self-attention layer, intermediate features serve as query and key that interacts with each other to achieve global consistency from different regions.

3.2. Text-to-layout generation

We aim to generate scene layouts that are both spatially well-organized and numerically accurate using LLMs such as GPT-4. Specifically, we provide user prompt, task description for layout generation and response example, then utilize LLMs' powerful visual concepts understand and numerical reasoning ability to support reasonable the positions and coordinates of objects according to the given response

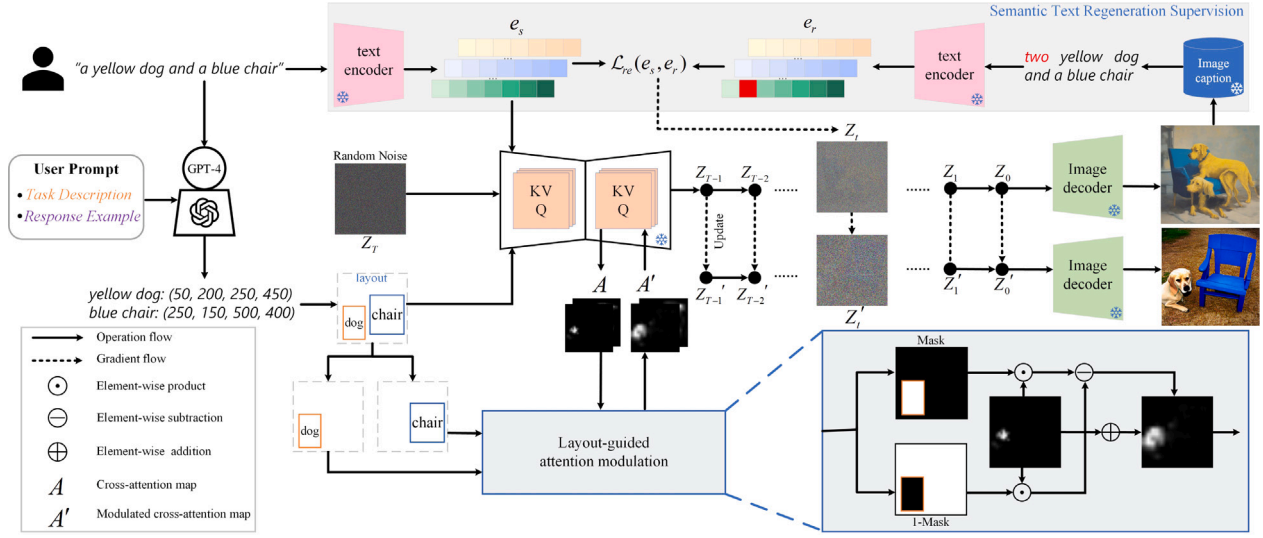

Fig. 3. The proposed MirrorDiff framework. The end-to-end architecture consists of three parts: Layout-Generation by LLM, Grounded Text-to-Image Generation, Layout-guided Attention modulation and Semantic Text Regeneration Supervision. At each denoising step, we modulate the cross-attention scores according to the mask of bounding box of main objects in order to ensure higher scores for major objects. After generating an image in each round, the noised sample is updated by optimizing semantic text regeneration supervision loss which aims to correct error regions and align synthetic images with given prompt in high-level semantic space.

Table 1
Our full prompt to the LLM for layout generation.

Role	Content
Instruction	You are an intelligent layout generator, your task is to assist users by providing helpful and precise coordinates information according to the given caption for an image. You should consider their sizes, positions and the number of objects. The size of the image is 512 × 512. The coordinates of the upper left corner are (0, 0) and the coordinates of the lower right corner are (512, 512). The format of each bounding box must be (object name, [coordinate of the top-left corner, coordinate of the top-left corner, width of the box, height of the box]) and it only includes one object. If necessary, you need to fully utilize your visual concept understand and layout reasoning ability. Please refer to the example below for the desired response format.
In-context examples	User: "Provide box coordinates for an image with one airplane on the top of four trucks." Assistant:"airplane: (50, 10, 460, 130), truck: (50, 180, 162, 350), truck: (178, 180, 290, 350), truck: (306, 180, 418, 350), truck: (434, 180, 506, 350)."
User prompt	User : "Provide the box coordinates for an image with" + [user prompt]

Table 2
The examples of prompts and responses with LLM.

Prompts: Provide box coordinates for an image with a red book and a green apple
Objects: red book: (100, 200, 250, 350), green apple: (300, 250, 400, 350)
Prompts: Provide box coordinates for an image with a green bench and a brown cat
Objects : green bench: (90, 160, 322, 312), brown cat: (340, 200, 452, 330)
Prompts: Provide box coordinates for an image with three elephants stand beside a pool of water
Objects : elephant 1: (30, 100, 200, 350), elephant 2: (210, 100, 380, 350), elephant 3: (390, 100, 560, 350), water pool: (150, 360, 420, 480)
Prompts: Provide box coordinates for an image with a horse and a car and the horse is larger than the car
Objects : horse: (25, 50, 450, 400), car: (250, 300, 435, 400)
Prompts: Provide box coordinates for an image with a airplane on the left of a horse
Objects : airplane: (12, 65, 275, 201), horse: (340, 135, 502, 368)
Prompts: Provide box coordinates for an image with a dog wearing a red hat is playing a guitar on the beach
Objects : dog: (150, 200, 350, 400), hat: (180, 150, 280, 200), guitar: (210, 250, 360, 380), beach: (0, 350, 512, 512)
Prompts: Provide box coordinates for an image with a box contains six donuts with varying types of glazes and toppings
Objects : box: (35, 35, 475, 475), donut1: (50, 50, 130, 130), donut2: (180, 50, 260, 130),donut3: (310, 50, 390, 130), donut4: (50, 200, 130, 280), donut5: (180, 200, 260, 280), donut6: (310, 200, 390, 280)

template. Similar designs with layoutGPT (Feng et al., 2024), given a text prompt condition C, the goal is to predict the positional coordinates $C = \{c_i | i = 1, 2, \ldots, n\}$ of objects $\mathcal{O} = \{o_i | i = 1, 2, \ldots, n\}$ in a series of texts, and $c_i = (l_i, t_i, r_i, b_i)$ indicates upper-left coordinate and lower-right coordinate of i-th object. Standardize response formats in samples to ensure that GPT is able to respond according to the standardize template which facilitates the extraction of coordinates from the responses to generate layouts. The details of full prompt and responses are shown in Table 1 and Table 2, respectively. Specifically, the full prompt mainly includes the three components:

- **Instruction** describes the task and specifies the response format to standardize the response content of LLM for convenient extraction of layout information.

- **In-context examples** serve to enhance the task understand ability of the model and are helpful for model to figure out the concepts of coordinates and layout.
- **User prompt** is added to the conversation to support format of user input.

3.3. Grounded text-to-image generation

Given a text prompt and a scene layout, grounded text-to-image generation model aims to synthesis a semantic and diverse image in specified locations. To further enhance the controllability and composability of grounded text-to-image model, we propose a semantic text regeneration supervision to correct error regions in the generation process by making the source text and regenerated text closer in high-level semantic space in order to align synthetic image with given text semantically. The latent z_t is updated by the gradient by computing loss between the source text with regenerated text. In addition, we propose a layout-guided attention modulation to reinforce attention score between token and interested region considering the size of the object to mitigate the disappearance of small objects. The framework of our proposed MirrorDiff is shown in Fig. 3.

3.4. Semantic text regeneration supervision

Facing a complex text containing multiple objects with multiple attributes and complicated spatial relationships, the problems of attribute coupling and quantity inaccuracy often occur when utilizing fundamental layout-guided text-to-image diffusion models. In order to alleviate it, we utilize pre-trained image caption model to regenerate text and propose a semantic text regeneration supervision to minimize the difference between the regenerated text and the given text in the semantic space to correct the wrong generated regions and attributes. As a result, the synthetic image aligns with the given text semantically.

During the experiment, we use frozen BLIP2 as image caption model to obtain the regenerated text. Inspired by leveraging CLIP text encoder to obtain text embedding into text-to-image diffusion model for image generation, we also introduce frozen CLIP text encoder to encode the given text and regenerated text into embedding space and achieve source text embedding e_s and regenerated text embedding e_r. As shown in top right corner of Fig. 3, the discrepancy between the generated image and the given text results in a divergence between the text embeddings of e_s and e_r, and the red regions in the embedding of the regenerated text embedding e_r indicate the erroneous generative elements implicitly. To correct it, we compute and minimize the semantic text regeneration supervision loss:

$$\mathcal{L}_{re}(e_s, e_r) = -log(cos(e_s, e_r)), \tag{3}$$

where e_s^i denotes the vector corresponding to the *i-th* token in source text embedding which is similar to e_r^i, M denotes the number of tokens, and $cos(*)$ calculates the cosine similarity. Having minimized the loss $\mathcal{L}_{re}$, we update the noised latent z_t at each denoising step as below :

$$z_t' = z_t - \alpha \bigtriangledown \mathcal{L}_{re}, \tag{4}$$

where α is the step size that balances the effect of the optimization in the denoising process which will be set to 2 for the first three rounds and subsequently reduced to 1.

3.5. Layout-guided attention modulation

Text and image features interact only in the cross-attention layer, and as the denoising process proceeds and t tends to 0, the distribution of the cross-attention map is similar to the given layout. At the same time, we found that the reason for the disappearance of small objects is that the token of small objects corresponds to a low attention score in the cross-attention map, while the attention score of large objects is high. To solve this problem, we propose layout-guided attention modulation presented in the bottom right corner of Fig. 3, which takes into account the size of the objects to modulate the cross-attention map and balance the cross-attention scores of small and large objects, and the process can be expressed as follows:

$$A' = softmax(\frac{QK^T + M}{\sqrt{d}}), \tag{5}$$

$$M_i = Mask(B_i) \odot QK^T \odot (1 - S_i) - (1 - Mask(B_i)) \odot QK^T \odot (1 - S_i), \tag{6}$$

where A' is the modulated cross attention map, $M \in \mathbb{R}^{|queries| \times |keys|}$ is the layout-guided attention modulation matrix, $Mask(B_i)$ is the binary mask generated from the bounding boxes B_i, S_i indicates the proportion of the area of the *i-th* bounding box region to the whole image, and $\odot$ is element-wise product. The detail of our algorithm can be summarized as **Algorithm** 1.

Algorithm 1: T2I Generation with LLMs by Redescription

Input: User Prompt P, Large Language Model ϕ, diffusion model θ, image caption model φ, clip encoder ρ, redescription round K, denoising step T.
Output: Synthetic image x
Obtain image layout positional coordinates $\mathcal{O} = \{o_j | j = 1, 2, ...n\}$ by Large Language Model $\phi(P)$
for $i = 1$ to K **do**
  for $t = T, T-1, ...1$ **do**
    for each cross attention map in θ **do**
      Extract prior layer's output z_i^t
      $Q_i^t = W_Q(z_i^t)$, $K_i^t = W_K(\rho(P))$
      Obtain cross attention map A_i^t from Q_i^t and K_i^t
      Obtain layout-guided attention modulation result $A_i^{t'}$ by A_i^t and Mask($\mathcal{O}$), and feed to following layers
    end for
    Generated image x_i by decoding(z_i^0)
    Obtain redescription $\varphi(x_i)$ by feeding x_i into image caption model
    Calculate semantic text regeneration loss L_{re} between $\rho(P)$ and $\rho(\varphi(x_i))$
    $z_i' \leftarrow z_i - \alpha_t L_{re}$
  end for
end for
return Generated image x_K

4. Experimental results and analysis

In this section, we first introduce the experimental setting of our method and present the qualitative and quantitative comparison results with state-of-the-art methods. Finally, we give the ablation studies and discussed the effects of each module and important hyperparameters.

4.1. Experimental settings

We choose the pre-trained GLIGEN (Li et al., 2023) as the base model and evaluate our method on several benchmarks. Futhermore, we also conduct ablation experiments for each component. All experiments are performed only on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory. All images are generated with 50 steps of denoising and the resolution of them are 512 × 512. In this paper, the Gaussian kernel size is 3 × 3 and the value of standard deviation σ is set 0.5. We add layout-guided attention modulation on all cross-attention maps of resolution 32 × 32, and the number of redescription iteration rounds K is 3.

Table 3
Quantitative comparisons with competing methods. On redescription iteration rounds K is set to 3 and choosing GPT-4.0 as a layout generator, we show the CLIP score and inference time on HRS, CC-500, NSR-1K and DrawBench. The highest and second scores in each section of the table are highlighted in **bold** and underlined respectively.

Model	HRS				CC-500	NSR-1K		DrawBench		Inference time (↓)
	Color	Spatial	Size	Counting		Spatial	Counting	Positional	Counting	
SD-v1.4	0.3241	0.3040	0.2898	0.3140	0.3186	0.3086	0.3033	0.2886	0.2085	**8.15 s**
SDXL	0.3383	0.3171	0.2976	0.3164	**0.3253**	0.3135	0.3031	**0.2969**	0.3297	8.97 s
Attend-and-Excite	0.3137	0.2982	0.2955	0.3143	0.3123	0.3021	0.2998	0.2915	0.3103	25.43 s
StructureDiffusion	0.3143	0.3023	0.2925	0.3139	0.3187	0.3028	0.3035	0.2882	0.3045	35.61 s
GLIGEN	0.3264	0.3083	0.2940	0.3072	0.3194	0.3052	0.2955	0.2897	0.3227	17.55 s
Ours ($K = 3$, GPT-4.0)	**0.3384**	**0.3237**	**0.3039**	**0.3171**	0.3249	**0.3164**	**0.3043**	0.2955	**0.3311**	21.83 s

4.1.1. Dataset

We utilize the HRS (Bakr et al., 2023), NSR-1K (Achiam et al., 2023) and DrawBench (Saharia et al., 2022) datasets to evaluate the performance of our method in compositional text-to-image generation task within multi-attribute, multi-object, and complex spatial relationships. In addition, in order to evaluate whether the model can accurately express the attribute information, we utilize the Concept Conjunction 500 (CC-500) (Feng et al., 2022) dataset. Specifically, the HRS dataset consists of four main categories: color, size, spatial and counting, and the number of prompts in each category is: 501/501/1002/3000. The NSR-1K dataset consists of 762 prompts about numerical relationships and 283 prompts about spatial relationships. The DrawBench dataset includes 39 prompts about counting and positional, and CC-500 dataset contains 600 text prompts approximately whose format is: "a [colorA] [objectA] and a [colorB] [objectB]".

4.1.2. Evaluation metrics

Following common practice (Jia and Tan, 2024; Phung et al., 2024), we choose to calculate object accuracy between images and text by YOLOv8 to evaluate the diversity and quality of the generated images. In particular, we also calculate precision, recall, F1 score, and accuracy in numeracy. The higher the above evaluation metrics, the better the quality of the generated images. In addition, to evaluate the similarity between text and image, we use CLIP score (Hessel et al., 2021) as evaluation metric. Furthermore, we calculate inference time to evaluate computational costs, which is beneficial to consider the trade-offs between performance and efficiency. What is more, we randomly extracted synthetic images on all categories of the HRS dataset to obtain human preferences by making a questionnaire.

4.2. Comparison with state-of-the-art methods

Baselines. We compare our methods with diffusion-based methods (e.g.: Stable Diffusion v1.4, Stable Diffusion XL, Attend-and-Excite and StructureDiffusion), layout-guided diffusion models (e.g.: Reco, GLIGEN and BoxDiff), and text-to-image diffusion models with LLMs (e.g.: LayoutGPT, Attention-and-Refocusing and DivCon), All the models mentioned above adopt the default settings from the original papers.

Quantitative results. Table 3 shows the results of quantitative comparison between our model and various state-of-the-art models on HRS, CC-500, NSR-1K datasets respectively. We achieve the highest clip-score on the HRS dataset and the NSR-1K dataset on all categories while outperforming the baselines. This enhancement can be attributed to semantic text regeneration supervision for correcting erroneous synthetic regions as well as to the effective modulation of the cross-attention map. In addition, on the CC-500 dataset, we rank second and are close to the best result. Moreover, our proposed MirrorDiff achieves a good balance between inference time and semantic alignment. The quantitative comparison with the text-to-image models using LLMs are shown in Table 4. On the CC-500 dataset, the performance of MirrorDiff is close to the best results, especially in the counting task, where our MirrorDiff outperforms LayoutGPT in terms of precision, F1 score and accuracy. As for spatial, we improve significantly compared with Attention-Refocusing (Phung et al., 2024).

Table 4
On redescription iteration rounds K is set to 3 and choosing GPT-3.5 or GPT-4.0 as a layout generator, we show the comparisons of our model with previous baselines with LLMs on the NSR-1K dataset. The best scores in each section are highlighted in **bold**.

Model	Couning				Spatial
	Precision↑	Recall↑	F1↑	Acc.↑	Acc.↑
LayoutGPT(GPT-3.5)	75.40	86.23	74.62	54.37	69.43
LayoutGPT(GPT-4.0)	81.02	85.63	78.11	56.25	70.96
Attention-Refocusing(GPT-4.0)	84.61	85.64	85.12	56.10	69.28
DivCon(GPT-4.0)	85.31	**87.54**	86.41	55.80	**72.95**
Ours($K = 3$, GPT-3.5)	78.62	86.93	82.44	55.31	69.97
Ours($K = 3$, GPT-4.0)	**85.37**	87.53	**86.43**	**56.79**	71.10

Qualitative results. Fig. 4 illustrates the visual comparison of our proposed model with other baselines including color attributes, spatial relationships, numeracy, and complex combinations. Without additional layout guidance, Stable diffusion (Podell et al., 2023), Structure Diffusion (Feng et al., 2022) and Attend-and-Excite (Chefer et al., 2023) have difficulty in generating spatial information that satisfies the semantic conditions. However, with layout guidance, the generated images by BoxDiff (Xie et al., 2023), GLIGEN (Li et al., 2023) and Reco (Yang et al., 2023) have poor image quality although the number of objects is basically consistent with the text prompt and fail to express the attributes of the objects correctly. For example, in the first row of Fig. 4, the input prompt is "A yellow dog and a blue chair", but the image generated by GLIGEN does not accurately express the colors of the dog and chair. On the contrary, our MirrorDiff not only correctly expresses the spatial relationship, but also generates the correct color attributes. Fig. 5 displays the additional visual comparison with competitive layout-guided text-to-image generation models including Reco, GLIGEN and BoxDiff which demonstrates our model outperforming in both simple and complex text. In a word, compared with the baselines, our proposed MirrorDiff is more effective regarding the fidelity to text content and layout conditions.

4.3. Ablation studies

In this section, we study the effects of different key components of our MirrorDiff quantitatively and qualitatively and conduct ablation experiments on the number of redescription round K. What is more, visualization results for the effect of LAM on Cross-Attention Map are also performed subsequently.

Effect of Key Components. We assess the key components on the HRS dataset and the quantitative evaluation results are shown in Table 5. We evaluated the effectiveness of semantic text regeneration supervision (STRS) and layout-guided Attention Modulation (LAM). The addition of STRS and LAM modules to GLIGEN models respectively and improves in all categories, especially in the color category, alleviates the attribute coupling phenomenon drastically.

Visualized ablation results are shown in Fig. 6. Only adding LAM to baseline enhances attentional scores within the bounding box but was insensitive to numeracy and attribute content. While adding STRS to baseline without LAM corrects erroneous attribute expressions, but lacks attention to the layout region. As a result, attribute information

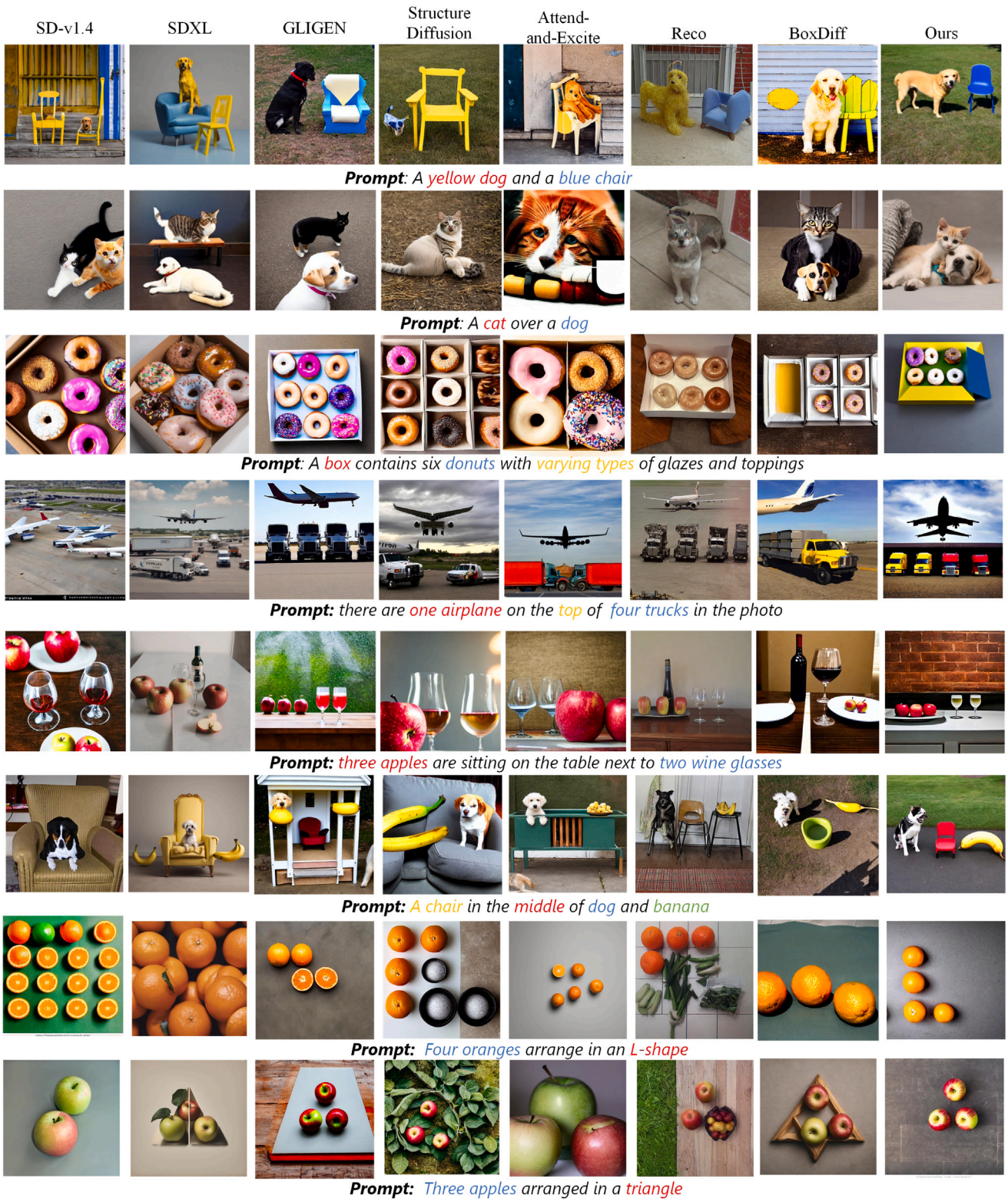

Fig. 4. Qualitative comparison between our approach and competing methods.

appears in the background in Fig. 6 (a) and residual representation of the object appears in the lower left corner outside the bounding box as shown in Fig. 6 (b) which is similar to the baseline.

Effect of the number of Redescription Round. The visual ablation results on Redescription Round are presented in Fig. 7 (a). As K gradually increases from 0 to 3, the generated images express attributes information more accurately. And when K=3, the generated images are basically consistent with the given text semantically. Therefore, K is set to 3 in all quantitative and qualitative experiments.

Effect of LAM on Cross-Attention Map. As shown in Fig. 7 (b), we visualize the cross-attention map of the main objects for a complex text. We found issues with object coupling and low attention scores for small objects, which is related to the causal attention masks of CLIP. While LAM emphasizes the expression of the bounding box region of

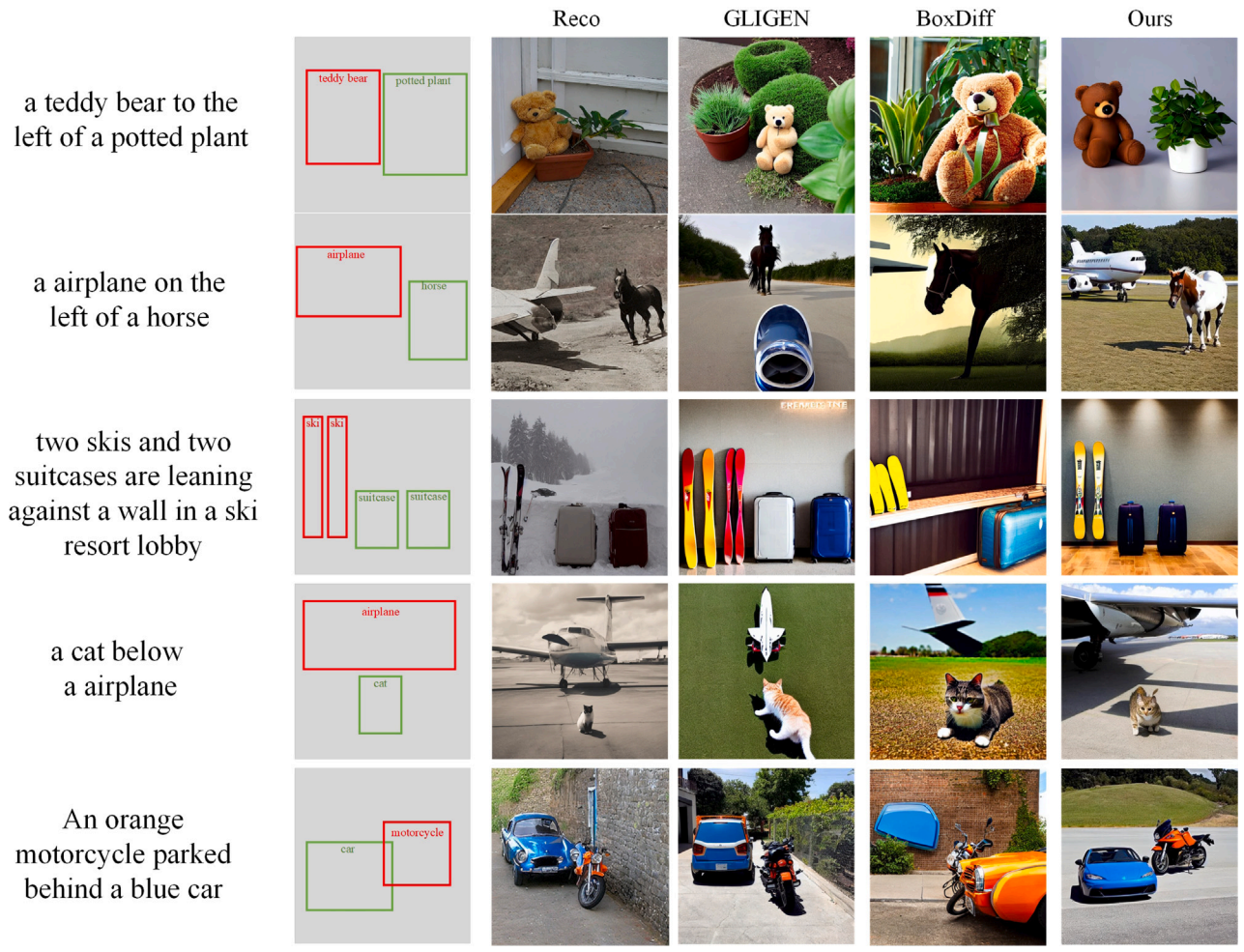

Fig. 5. Visual comparisons of ours and other layout-guided text-to-image generation models.

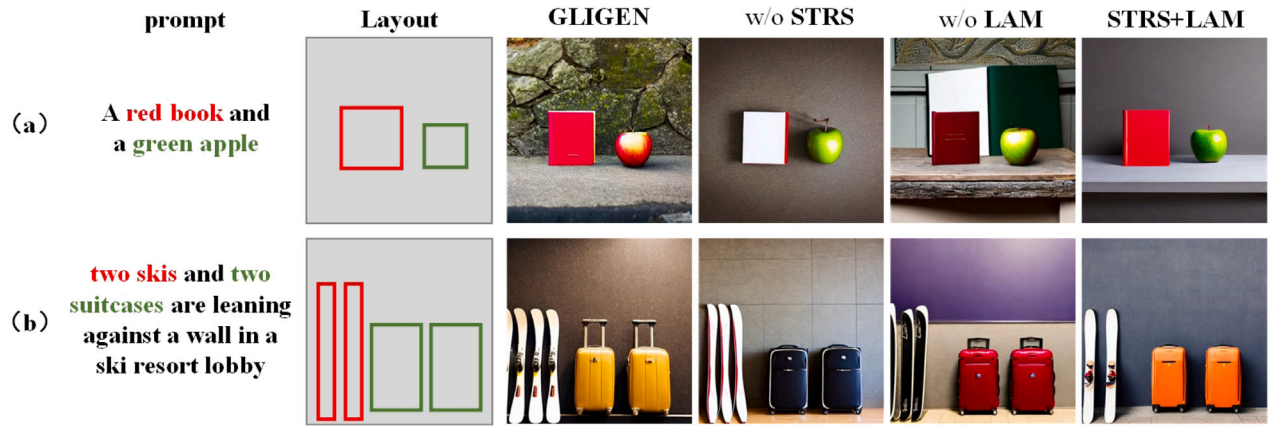

Fig. 6. An visual ablations analysis of the Impact of various components on MirrorDiff. The layout obtained by LLMs guidelines are delineated on the images using dashed boxes.

the token and mitigates the above-mentioned issues, especially boosting the attention score for small objects.

Comparison between GPT-3.5 and GPT-4.0. The comparison between GPT-3.5 and GPT-4.0 in layout generation is presented in Fig. 8. Compared with GPT-3.5, GPT-4.0 can accomplish the reasonable layout generation task well. Specifically, in the last prompt example, GPT-4.0 extracts main objects in detail and supports more precision layout, unlike GPT-3.5 which describes the red hat and the dog in one bounding box. Therefore, we use GPT-4.0 to help users generate layouts in our experiments.

4.4. Additional results

In this section, we investigate the impact of dynamic prompt engineering on MirrorDiff's performance. Dynamic prompt engineering involves modifying the linguistic styles, complexity, and object descriptions in the text prompts to explore how these variations affect the

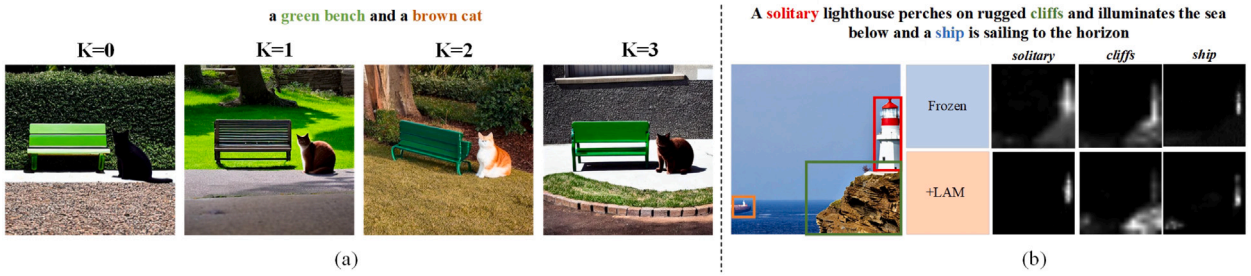

Fig. 7. Visual ablations analysis on redescription round and impact of LAM on cross-Attention map.

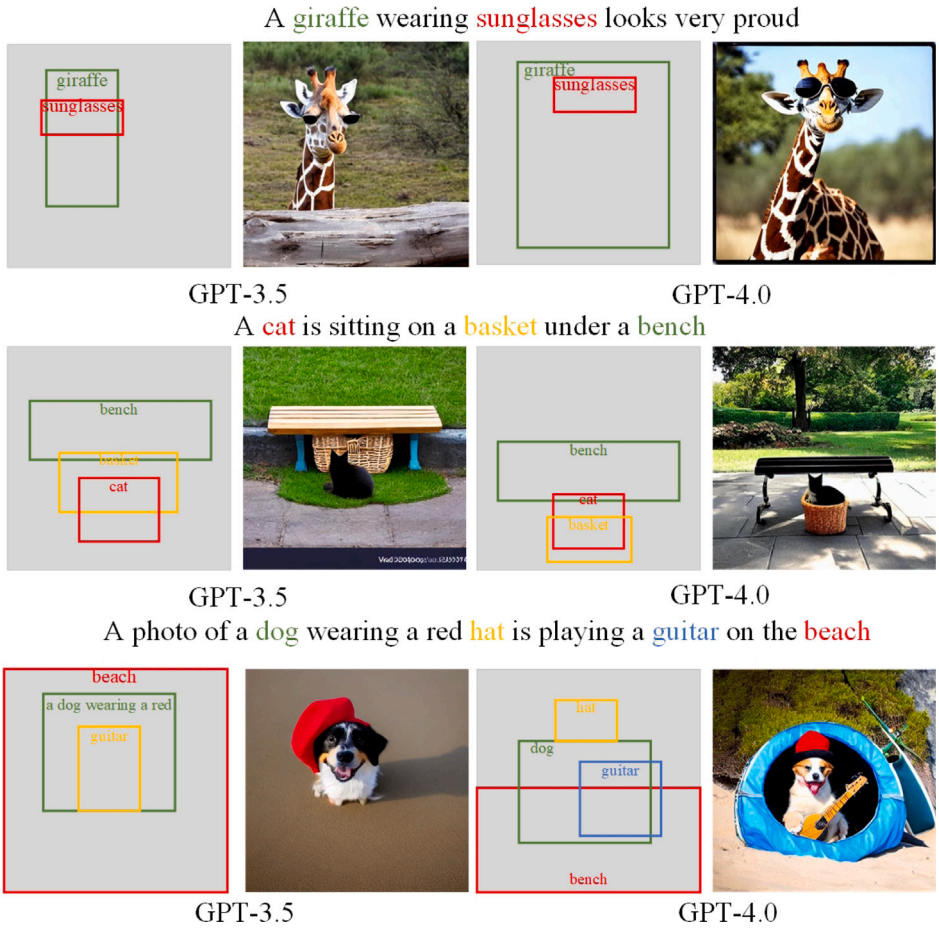

Fig. 8. Comparison generated layouts from GPT-3.5 and GPT-4 and synthetic images by our proposed MirrorDiff.

Table 5
Abaltion studies of the STRS and LAM using the GLIGEN model on the HRS benchmark.

STRS	LAM	HRS			
		Color	Spatial	Size	Counting
×	×	0.3242	0.3140	0.2903	0.3127
×	√	0.3298	0.3199	0.2994	0.3143
√	×	0.3341	0.3203	0.3001	0.3152
√	√	**0.3384**	**0.3237**	**0.3039**	**0.3171**

performance of MirrorDiff, which is crucial because the quality and specificity of the text prompt significantly influence the accuracy and controllability of text-to-image generation models. As shown in Table 6, we design three experimental plans for varying language styles, gradually increasing complexity and different object descriptions, respectively. The visualization results of three experimental plans are performed in Fig. 9, which demonstrate that our MirrorDiff can adapt to dynamic input prompts and perform well in terms of semantic alignment and spatial controllability.

5. Limitation and discussion

Although our MirrorDiff performs well in most cases, but still remains a few limitations. **On the one hand,** the responses of GPT are sometimes nonstandard and uncontrollable. Despite providing response template in the in-context examples to ensure that the responses of GPT are predominantly accurate in the majority of scenarios, the emergence of unexpected response occurs inevitably. In the future, we plan to further explore the capabilities of LLMs in concept understanding and layout planning by designing finer-grained templates and constraint mechanisms to improve the stability and controllability of their responses. In addition, we will investigate how to incorporate other advanced LLM techniques, such as fine-tuning or domain-specific adaptive training, to further enhance the quality of layout generation.

Table 6
Experiment settings for investigating dynamic prompt engineering on MirrorDiff.

Experiment plans	Type	Input prompt
Experiment-1 (varying prompt styles)	Normal style Art style	A red ball on the left side of a blue cube A crimson orb rests beside a cerulean block, bathed in the golden light of the setting sun
Experiment-2 (complexity)	Low complexity Medium complexity High complexity	A red ball A red ball and a blue cube A yellow chair in the middle of a red ball and a blue cube
Experiment-3 (object descriptions)	Vague description Detailed description	A car on the road A red sports car with black rims speeding on a winding mountain road

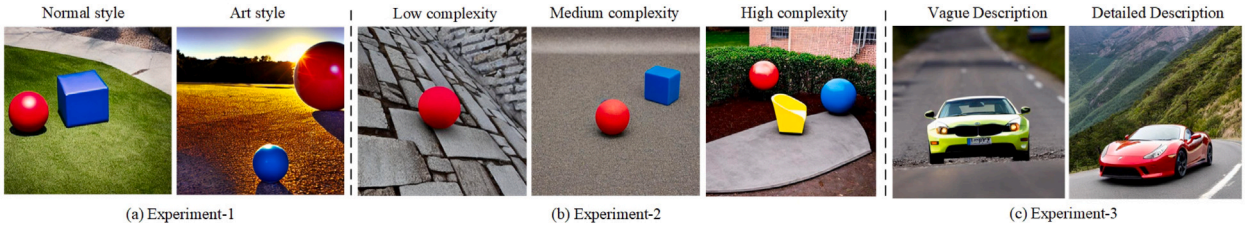

Fig. 9. Visualization results of dynamic prompt engineering on MirrorDiff's performance by varying linguistic styles, complexity, and object descriptions.

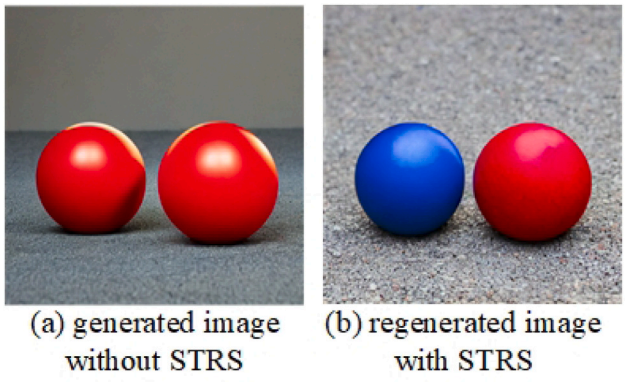

Fig. 10. Failure case. (a) the generated image without STRS, (b) regenerated image with STRS. Although STRS corrects color errors, fails to correct position and size errors.

On the other hand, STRS performs well in most cases, but it may not be able to fully correct all discrepancies because the pre-trained image caption model sometimes fails to accurately express image information and ignores some attribute representations. For instance, the original text is "a large red ball on left and a small blue ball on the right", but the size and color of the two balls in the generated image are not accurate (in Fig. 10 (a)). While the regenerated text is "two red balls", losing the details of size and position, and only corrects wrong color attribute. The image corrected by STRS expresses the correct color information (in Fig. 10 (b)), but due to the loss of size and position, fails to correct the size and position of the image. Additionally, we tested the probability of failing cases due to STRS on the CC-500 dataset, and the failure rates were 4.7%, 2.6%, and 2.1% under the number of redescription round is 1, 2 and 3, which indicates that our proposed STRS can work in most cases.

In this paper, our proposed MirrorDiff addresses several critical limitations in existing text-to-image generation models, particularly in handling complex prompts with multiple objects and attributes. Specifically, MirrorDiff leverages LLM for layout generation and introduces a layout-guided attention modulation strategy to modulate attention maps adaptively for increasing the attention of main objects. Additionally, in contrast to existing methods that can only generate images but fail to correct erroneous generated regions, we propose semantic text regeneration supervision to ensure semantic consistency between the generated image and the original text by reducing the discrepancy between the original text and the regenerated text, which can further correct erroneous content generation iteratively. Currently, our MirrorDiff focuses on text-to-image generation tasks, lacking multimodal interaction (e.g.: text, image and speech). In future research, we plan to extend the MirrorDiff to multimodal generation tasks, such as generating images based on speech input, which enables our model to handle more complex multimodal tasks, further enhancing its application value.

6. Conclusion

In this paper, we propose a novel training-free grounded text-to-image-to-text framework to enhance the alignment of synthetic images with given text and correct error regions by prompt redescription. Specifically, we design a layout-guided attention modulation strategy to adjust attention map effectively to mitigate the disappearance of small objects. In addition, a semantic text regeneration supervision is devised to constrain regenerated text and the original text align in high-level semantic space to correct erroneous attribute and spatial relationship expressions of generated image. Extensive experiments show that our model outperforms state-of-the-art compositional grounded text-to-image models.

CRediT authorship contribution statement

Chang Liu: Writing – original draft, Software, Methodology, Data curation. **Mingwen Shao:** Writing – review & editing, Writing – original draft, Conceptualization. **Zhengyi Gong:** Writing – original draft, Software, Methodology, Data curation. **Xiang Lv:** Writing – original draft, Software, Methodology, Data curation. **Lingzhuang Meng:** Writing – original draft, Software, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFA1000102), the National Natural Science Foundation of China (Grant Nos. 62376285 and 61673396), and Natural Science Foundation of Shandong Province, China (Grant No. ZR2022MF260).

Data availability

Data will be made available on request.

References

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

Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X., 2023. Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18370–18380.

Bakr, E.M., Sun, P., Shen, X., Khan, F.F., Li, L.E., Elhoseiny, M., 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20041–20053.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al., 2023. Improving image generation with better captions. Comput. Sci. https://cdn.openai.com/papers/dall-e-3.pdf 2, 8.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al., 2023. Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D., 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph. 42, 1–10.

Chen, M., Laina, I., Vedaldi, A., 2024. Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353.

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

Couairon, G., Careil, M., Cord, M., Lathuiliere, S., Verbeek, J., 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2174–2183.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al., 2021. Cogview: Mastering text-to-image generation via transformers. Adv. Neural Inf. Process. Syst. 34, 19822–19835.

Ding, M., Zheng, W., Hong, W., Tang, J., 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. Adv. Neural Inf. Process. Syst. 35, 16890–16902.

Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y., 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032.

Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y., 2024. Layoutgpt: Compositional visual planning and generation with large language models. Adv. Neural Inf. Process. Syst. 36.

Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y., 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718.

Jia, Y., Tan, W., 2024. Divcon: Divide and conquer for progressive text-to-image generation. arXiv preprint arXiv:2403.06400.

Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T., 2023. Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134.

Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y., 2023. Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J., 2023. Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521.

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

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., 2024. 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. pp. 4296–4304.

Nie, W., Liu, S., Mardani, M., Liu, C., Eckart, B., Vahdat, A., 2024. Compositional text-to-image generation with dense blob representations. arXiv preprint arXiv:2405.08246.

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

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R., 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR. pp. 8748–8763.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1, 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. 35, 36479–36494.

Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C., 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16515–16525.

Wang, R., Xiong, J., Ke, H., Jia, Y., Wang, D.D., 2023. Improving the adversarial robustness of deep neural networks via efficient two-stage training. In: 2023 International Conference on Machine Learning and Cybernetics. ICMLC, IEEE, pp. 43–49.

Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z., 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461.

Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al., 2023. Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255.

Yang, B., Xiang, X., Kong, W., Zhang, J., Peng, Y., 2024. Dmf-gan: Deep multimodal fusion generative adversarial networks for text-to-image synthesis. IEEE Trans. Multimed..

Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al., 2022. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2, 5.

Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y., 2021. Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 833–842.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.