Text-to-Image Generation Method Based on Object Enhancement and Attention Maps

Yongsen Huang, Xiaodong Cai, Yuefan An
School of Information and Communication, Guilin University of Electronic Technology, Guilin, China

***Abstract*—In the task of text-to-image generation, common issues such as missing objects in the generated images often arise due to the model's insufficient learning of multi-object category information and the lack of consistency between the text prompts and the generated image contents. To address these challenges, this paper proposes a novel text-to-image generation approach based on object enhancement and attention maps. First, a new object enhancement strategy is introduced to improve the model's capacity to capture object-level features. The core idea is to generate difficult samples by processing the object mask maps of tokens, followed by dynamic weighting of the attention map using latent image embeddings. Second, to enhance the consistency between the text prompts and the generated image contents, we enforce similarity constraints between the cross-attention maps and the attention-weighted mask feature maps, penalizing inconsistencies through a loss function. Experimental results demonstrate that the Stable Diffusion v1.4 model, optimized using the proposed method, achieves significant improvements on the COCO instance dataset and the ADE20K instance dataset. Specifically, the MG metrics are improved by an average of 12.36% and 6.55%, respectively, compared to state-of-the-art models. Furthermore, the FID metrics show a 0.84% improvement over the state-of-the-art model on the COCO instance validation set.**

***Keywords*—Multi-object category; text-to-image generation; object enhancement; attention maps**

I. Introduction

The rapid advancement of artificial intelligence has propelled text-to-image generation into the forefront of research at the intersection of computer vision and natural language processing. This emerging field aims to automatically generate visual content that aligns with natural language descriptions. Beyond its theoretical significance, this technology holds vast potential for application in diverse areas, including virtual reality, game design, artistic creation, and human-computer interaction. However, a major challenge persists: efficiently translating textual semantics into high-quality images while ensuring a high degree of consistency between the generated images and their corresponding textual descriptions.

Recent advances in deep learning have opened new avenues for text-to-image generation tasks. Generative Adversarial Networks (GANs) [1], as an early foundational technology, enabled the initial mapping from text to image via adversarial training between generators and discriminators. However, the quality and semantic alignment of the generated images still require significant improvement. The subsequent development of autoregressive and diffusion models has invigorated the field. Autoregressive models generate high-quality, semantically consistent images through pixel-by-pixel synthesis but suffer from slow training and inference speeds. In contrast, diffusion models generate images through a process of iterative denoising, yielding not only high-quality images but also enhanced diversity, positioning them as the prevailing approach in contemporary research.

The objective of diffusion models can be summarized as reversing the gradual degradation process of data, which consists of a forward process that follows a Markov chain and a reverse diffusion process. In the forward process, noise is gradually added to the original data, causing it to degrade into nearly isotropic Gaussian noise, thereby corrupting the original data. In contrast, the reverse diffusion process utilizes a neural network to learn how to recover the original data from Gaussian noise. It is important to note that the input and output dimensions of the reverse diffusion process must remain consistent.

Although recent text-to-image diffusion models [2][3][4][5][6][7][8][9] have achieved notable progress in generating images with increasing levels of quality, resolution, realism, and diversity, a significant challenge remains in maintaining consistency between text prompts and the content of the generated images.

Several studies have addressed the challenge of generating images containing multiple objects. In 2022, Robin Rombach et al. [10] introduced Stable Diffusion, a high-resolution image synthesis method based on Latent Diffusion Models (LDMs), designed to overcome the computational inefficiencies of traditional diffusion models in high-resolution image generation. The model achieves diverse high-resolution image generation by integrating components such as variational autoencoders (VAE), conditional text encoders, and U-Net. In the same year, Liu et al. [11] proposed Composable Diffusion, a method leveraging multiple diffusion models to generate complex scenes by separately generating different objects, each handled by a specialized model. That same year, Liew et al. [12] introduced MagicMix, a technique that uses pre-trained, text-conditioned diffusion models to blend two distinct semantic concepts into a single image. The process begins by generating a rough semantic layout, followed by content matching the text description, and concludes by merging the semantic information of the two objects. In 2023, Chefer et al. [13] developed Attend-and-Excite, a method aimed at enhancing the semantic fidelity of text-to-image diffusion models. By optimizing the cross-attention mechanism, this approach ensures that the generated images better reflect the input text prompts. Directed Diffusion [14], also in 2023, introduced a novel approach to controlling the positioning of multiple elements in the image by manipulating attention maps at the word and word-position

Guangxi Driven Development Project (桂科 AA20302001).

levels, improving the model's focus on the relevant areas of the image. In the same year, Zirui Wang et al. [15] presented TokenCompose, a method that incorporates token-level supervision to improve performance in multi-object composition tasks and enhances the photorealism of generated images. More recently, in 2024, Tobias Lingenberg et al. [16] proposed DIAGen, an image enhancement method for few-shot learning scenarios. By combining generative models with text prompts, the method effectively increases the diversity of generated images.

However, most prior research has been limited to simple augmentation techniques, such as flipping, rotation, or basic enhancement operations on image data. These methods often fail to adequately capture the object features, leading to issues such as missing objects in the generated images.

Moreover, while much of the previous research has focused on the spatial layout of the cross-attention maps between text prompts and generated image contents during image generation, it has often overlooked the importance of enhancing the understanding of the spatial layout of the object cross-attention maps during the model's training phase.

To address the aforementioned issues, inspired by the literature [10] [15] , this paper proposes a novel text-to-image generation method based on object enhancement and attention maps (TI-OEAM). By constructing difficult samples, incorporating a dynamic residual gating mechanism, and applying an attention maps guidance approach, this paper optimizes the model and significantly enhances the consistency between the text prompts and the generated images, leading to a substantial improvement in image quality.

The subsequent sections of this paper will provide a detailed exploration of this research. The TI-OEAM model section will describe the proposed method, focusing on the design and implementation of the object enhancement strategy, as well as how attention map optimization is employed to improve the consistency between text prompts and the generated image content. The experimental section will outline a series of experiments conducted to evaluate the effectiveness and performance of the proposed approach, including comparative studies with existing methods and ablation experiments. Finally, the conclusion section will summarize the key findings and contributions of this work, discuss its limitations, and propose directions for future research.

II. The Proposed Framework

A. An Overview of the Proposed Framework

As shown in Fig. 1, in the object enhancement module, the similarity between the object mask map and all other object mask maps in the image is first evaluated through a similarity calculation. Object mask maps with similarity values exceeding a predefined threshold are then filtered out. Gaussian noise is then applied to these selected mask maps to introduce random perturbations, generating difficult samples. Subsequently, latent noisy image embeddings with learnable parameters are utilized as dynamic residual weighting terms in the denoising U-Net, which are integrated into the model's training process. Building on this, attention map optimization is performed by imposing similarity constraints between the cross-attention map and the attention-weighted masked feature map. Finally, the model is optimized using a loss function.

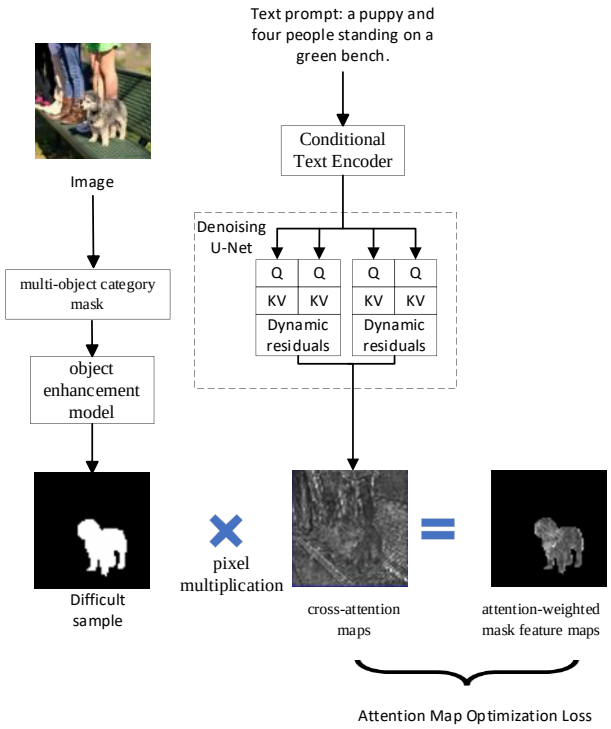

Fig. 1. TI-OEAM overall framework.

B. Object Enhancement Strategies using Masking and Residuals

To fully capture the diverse object feature information in images, a novel object enhancement strategy is proposed, consisting of two key steps. The first step involves constructing difficult samples, while the second step introduces a dynamic residual gating mechanism.

1) Difficult sample generation: In an image, multiple objects often exhibit similar visual characteristics, particularly when they belong to the same category or share similar appearances. Such similarities may hinder the model's ability to accurately distinguish between objects, potentially resulting in missed or incorrect generation of specific objects during image generation. To address this issue, we propose calculating the similarity between object masks and introducing noisy interference to highly similar masks. This approach forces the model to focus more on the subtle differences between similar objects. The interference thus encourages the model to learn how to differentiate and generate diverse features of similar objects, ultimately enhancing its capacity to generate objects across multiple categories more effectively.

Specifically, the first step involves designing an object enhancement module that calculates the pixel-level feature similarity between the object mask map in the image and all other object mask maps. Object mask maps with similarity values exceeding a predefined threshold are then filtered out. Subsequently, Gaussian random interference is applied to these mask maps to generate difficult samples.

The specific calculation process is as follows:

$$f = \frac{1}{G\text{-}1}\sum_{k}^{G\text{-}1}\cos(\boldsymbol{M}, \boldsymbol{M}_k) \tag{1}$$

Where $\boldsymbol{M}$ denotes the object mask map in the image, G denotes the number of object mask maps the image contains,and f indicates the average pixel-level feature similarity between the object mask map and all other object mask maps in the image.

Then Gaussian random interference is applied to the object mask maps above the set threshold in the following process:

$$\begin{cases} \boldsymbol{M}' = \boldsymbol{M} + \Delta', f > \varphi \\ \boldsymbol{M}, f \le \varphi \end{cases} \tag{2}$$

Where the threshold φ is set to 0.8, the added noise vector Δ' obeys $\| \Delta \|_2 = \grave{o}$,and $\grave{o}$ is a small constant. $\boldsymbol{M}'$ is called the difficult sample, where Δ process is as follows:

$$\Delta = \omega, \omega \in R^d \square U(0,1e-3) \tag{3}$$

The incorporation of difficult samples encourages the model to focus on object mask maps that exhibit high similarity and are challenging to distinguish from other objects in the image. This strategy enables the model to more effectively learn the distinctive features of each object.

Dynamic Residual Gating: In this study, the definition proposed by Zirui Wang et al. [15] is adopted. For a given image $x \in R^{H\times W\times 3}$ in RGB space， it is first processed by the encoder part of the VAE to obtain its latent image embedding $\mathrm{z}_0 = \varepsilon(x_0)$. Subsequently, based on this potential image embedding, Gaussian random noise is injected with time t to obtain the potential noisy image embedding z_t containing the noise. Additionally, a conditional text encoder is employed to convert the text prompts y into an embedding $\tau_\theta(\mathrm{y})$ with the neural network parameter θ .Let $\boldsymbol{K} \in R^{H\times L_{\tau_\theta(y)}\times d_k}$ represent the embedding of the text prompts corresponding to the token， where $L_{\tau_\theta(y)}$ denotes the length of the text prompts embedding $\tau_\theta(\mathrm{y})$ .Let $\boldsymbol{Q} \in R^{H\times L_{z_t}\times d_k}$ represent the latent noisy image embedding, with L_{z_t} indicating the length of the latent noisy image embedding. d_k denotes the dimension of $\boldsymbol{K}$ .The cross-attention map of text prompts and images is computed as follows:

$$\boldsymbol{Q}^{(h)} = \boldsymbol{W}_Q^{(h)} \cdot \varphi(\mathrm{z}_t) \tag{4}$$

$$\boldsymbol{K}^{(h)} = \boldsymbol{W}_K^{(h)} \cdot \tau_\theta(y) \tag{5}$$

$$\boldsymbol{V}^{(h)} = \boldsymbol{W}_V^{(h)} \cdot \tau_\theta(y) \tag{6}$$

$$\boldsymbol{A} = \frac{1}{H}\sum_{h}^{H}\mathrm{soft\,max}(\frac{\boldsymbol{Q}^{(h)}(\boldsymbol{K}^{(h)})^T}{\sqrt{d_k}})\cdot\boldsymbol{V} \tag{7}$$

Where $h \in \{1,...,H\}$ denotes each head in the multi-head cross-attention. $\boldsymbol{W}_Q^{(h)}$, $\boldsymbol{W}_K^{(h)}$ and $\boldsymbol{W}_V^{(h)}$ are the learnable projection matrices in the cross-attention layer in each head. $\varphi(\mathrm{z}_t)$ denotes the function that flattens the 2D latent image embedding to 1D.

This study identifies a limitation in most current approaches, which directly use the output of the pre-trained U-Net model as the input for the VAE-generated images. This practice often results in insufficient learning of multi-category object information. To address this, the second step involves designing a dynamic residual gating mechanism.

In this paper, learnable parameters are employed to adjust the weights of the residual connections, thereby allowing the model to flexibly control the flow of information based on varying contextual conditions. This dynamic adjustment mechanism enhances the model's nonlinear learning capability, enabling more accurate representation of image features. By adopting this approach, the model can optimize the learning process during image generation in accordance with the specific characteristics of the image, thus mitigating the issue of insufficient feature capture that may arise when the weights of residual connections are fixed.

Specifically, the latent image embedding is used to preserve image features, thereby allowing the model to concentrate more effectively on these features. The residual cross-attention computation process is as follows:

$$\boldsymbol{A}' = \boldsymbol{A} + \alpha * \mathrm{z}_t \tag{8}$$

Where α is the learnable parameter. The dynamic residual gating mechanism enhances and refines the cross-attention maps by dynamically adjusting the weights of the residual connections. This mechanism improves the model's ability to capture features, enabling it to more accurately learn the features of the data.

C. Optimization of Attention Map

Robin Rombach et al. [10] and Zirui Wang et al. [15] learned noisy features by using a denoising function, L_{DM} , which takes into account the lack of direct optimization correlation between tokens and image contents. To compensate for this deficiency, they introduced a token-level attention loss, L_t , which monitors the activation region of cross-attention. In addition, to prevent attention from being overly focused on certain sub-regions of the target region, they also propose a pixel-level attention loss L_p .

Inspired by the work of Robin Rombach, Zirui Wang, and others, we aim to better manage the spatial arrangement and interactions of multiple objects during the image generation process, ensuring that the generated image accurately includes

all specified categories and objects. In this paper, we propose a novel attention optimization method that constrains the similarity between the cross-attention map and the attention-weighted masked feature map. The objective is to improve the model's sensitivity to information within specific attention regions, without compromising its ability to capture global context. Let i denote the noun toekn of a text prompt. Let $\boldsymbol{A}_i'$ represent the cross-attention map between the latent noisy image embeddings and the embedding of a token.

First, the attention-weighted mask feature map is computed as follows:

$$\boldsymbol{A}'_{(i,u)} \square \boldsymbol{M}'_{(i,u)} = \boldsymbol{A}'_{M'(i,u)} \tag{9}$$

Where $\square$ denotes the pixel multiplication . u is the spatial location of the cross-attention map. $\boldsymbol{M}'_{(i,u)}$ denotes the object mask map of text token i at spatial position u . Let $\boldsymbol{A}'_{(i,u)}$ denote the cross-attention map formed by the latent noisy image embedding of the i -th token at the spatial location u of $\boldsymbol{A}_i' \in R^{L_{z_t}}$.

Subsequently, the cosine similarity function is calculated in accordance with the following procedure:

$$\mathrm{S}(\boldsymbol{A}'_M, \boldsymbol{A}') = \frac{\boldsymbol{A}'_{M'(i,u)} \cdot \boldsymbol{A}'_{(i,u)}}{|\boldsymbol{A}'_{M'(i,u)}| \cdot |\boldsymbol{A}'_{(i,u)}|} = \sum_i^{L_{z_t(r)}} \sum_u^{L_{z_t}} \frac{\boldsymbol{A}'_{M'(i,u)} \boldsymbol{A}'_{(i,u)}}{\sqrt{\boldsymbol{A}'^{\,2}_{M'(i,u)}} \sqrt{\boldsymbol{A}'^{\,2}_{(i,u)}}} \tag{10}$$

Where $\mathrm{S}(\boldsymbol{A}'_{M'}, \boldsymbol{A}')$ denotes the similarity between the cross-attention map and the attention-weighted mask feature map.

The two are then brought into closer alignment in the pixel domain using a loss function, which is calculated as follows:

$$\boldsymbol{L}_s = 1 - \mathrm{S}(\boldsymbol{A}'_{M'}, \boldsymbol{A}') \tag{11}$$

Cosine similarity function $\mathrm{S}(\boldsymbol{A}'_{M'}, \boldsymbol{A}')$ measures how close the cross attention map is to the attention-weighted masked feature map in feature space. The loss function L_s effectively balances the learning of both local and global information, overcoming the limitations of traditional methods that tend to over-focus on the target region or neglect other areas of the image. This balance improves the alignment between text prompts and generated image content, thereby enhancing both image consistency and overall quality.

D. Loss Function

Finally, L_s and L_{DM}, L_t and L_p jointly trained and computed as follows:

$$L = L_{DM} + \sum_d^D \left(\gamma L_s + \eta L_t + \psi L_p \right) \tag{12}$$

Where η , γ and ψ are the scaling factors. In this paper, we set γ as 1e-3, η as 1e-3, ψ as 5e-5。 The value of D represents the number of training layers.

III. Experiment

A. Datasets

In order to evaluate the effectiveness of the proposed model in the text-to-image generation task, this paper utilizes the COCO dataset [15] for training experiments. This dataset consists of 4,526 image-caption pairs and their corresponding binary mask maps. The choice of COCO is motivated by the relatively low ambiguity in its visual language and the rich diversity of object categories represented in each image, which effectively supports the task of learning and generating multi-category objects. To further assess the model's performance in multi-category instance combination, we conduct experiments on the COCO instance dataset [17], which includes 80 categories, and the ADE20K instance dataset [18][19], which includes 100 categories. Additionally, to evaluate the distributional differences between the images generated by the model and real images, we randomly sampled 10,000 image-caption pairs from the COCO instance validation set (C) and 1,000 image-caption pairs from the Flickr30K instance validation set (F) [20] for comparative analysis.

B. Evaluation Metrics

In this paper, we use the MULTIGEN [15] metric and the FID [21] metric to assess the ability of the model in combining instances of multiple categories and to analyse the distributional differences between the images generated by the model and the real images.

MULTIGEN is a challenging metric used to evaluate multi-category instance combinations. Specifically, given a set of N distinct instance categories, five categories (e.g., A, B, C, D, and E) are randomly selected and formatted into a sentence (e.g., "A photo of A, B, C, D, and E"). This sentence serves as the conditional input for the text-to-image diffusion model to generate the corresponding image. Subsequently, a robust open-vocabulary detector [15] is employed to assess whether the specified categories are accurately represented in the generated image.

Specifically, for each dataset, 1,000 text prompts were generated by randomly sampling 1,000 instances from each of the 80 COCO categories and 100 ADE20K categories, which were then used as inputs for the multi-category instance combinations. For each text prompt, 10 rounds of image generation were performed, resulting in a total of 10 × 1000 images for each dataset's category combination. Each generated image was subsequently analyzed using a detector to count the number of category instances present. Based on these detection results, the MG2 to MG5 metrics were computed for each round.

The mean and standard deviation (denoted in parentheses) of the MG2-5 success rates across the 10 rounds are presented in Table I.

TABLE I. COMPARISON OF EXPERIMENTAL RESULT OF VARIOUS MODELS

Method	Multi-category Instance Composition↑								Photorealism↓	
	COCO INSTANCES				ADE20K INSTANCES				FID (C)	FID (F)
	MG2	MG3	MG4	MG5	MG2	MG3	MG4	MG5		
SD	$90.72_{1.33}$	$50.74_{0.89}$	$11.68_{0.45}$	$0.88_{0.21}$	$89.81_{0.40}$	$53.96_{1.14}$	$16.52_{1.13}$	$1.89_{0.34}$	20.88	71.46
Composable	$63.33_{0.59}$	$21.87_{1.01}$	$3.25_{0.45}$	$0.23_{0.18}$	$69.61_{0.99}$	$29.96_{0.84}$	$6.89_{0.38}$	$0.73_{0.22}$	-	75.57
Layout	$93.22_{0.69}$	$60.15_{1.58}$	$19.49_{0.88}$	$2.27_{0.44}$	$96.05_{0.34}$	$67.83_{0.90}$	$21.93_{1.34}$	$2.35_{0.41}$	-	74.00
Structured	$90.40_{1.06}$	$48.64_{1.32}$	$10.71_{0.92}$	$0.68_{0.25}$	$89.25_{0.72}$	$53.05_{1.20}$	$15.76_{0.86}$	$1.74_{0.49}$	21.13	71.68
Attn-Exct	$93.64_{0.76}$	$65.10_{1.24}$	$28.01_{0.90}$	**$6.01_{0.61}$**	$91.74_{0.49}$	$62.51_{0.94}$	$26.12_{0.78}$	$5.89_{0.40}$	-	71.68
TokenCompose	$98.08_{0.40}$	$76.16_{1.04}$	$28.81_{0.95}$	$3.28_{0.48}$	$97.75_{0.34}$	$76.93_{1.09}$	$33.92_{1.47}$	$6.21_{0.62}$	20.19	**71.13**
TI-OEAM	**$98.54_{0.19}$**	**$79.43_{0.91}$**	**$32.55_{0.96}$**	$4.32_{0.36}$	**$97.91_{0.18}$**	**$79.39_{0.85}$**	**$37.13_{2.39}$**	**$7.04_{0.51}$**	**20.02**	71.94

The FID metric is used to assess the distributional disparity between two datasets: 10,000 image-caption pairs from the COCO instance validation set (C) and 1,000 image-caption pairs from the Flickr30K instance validation set (F), in comparison with the model-generated images.

C. Experimental Settings

The experimental setup is as follows: The operating system is Ubuntu 18.04.5 LTS; the hardware configuration includes an NVIDIA 3090 GPU; the deep learning framework used is PyTorch; and the programming language is Python.

The primary experiments in this paper are conducted using the Stable Diffusion v1.4 [22] model and the TokenCompose model. Stable Diffusion is a widely used text-to-image diffusion model for high-quality generation. AdamW is employed as the optimizer [23], with a global learning rate of 5e-6 and a total of 2400 steps. The image resolution is set to 512. Training was performed on a single GPU using a single batch and four gradient accumulation steps across the entire U-Net.

D. Experimental Results and Analysis

To validate the effectiveness of the TI-OEAM model, this study compares its performance with several representative text-to-image generation methods. These include Stable Diffusion (SD) [22], which addresses the high computational cost of traditional diffusion models in high-resolution image generation; Composable Diffusion [11], which generates complex scenes by combining multiple diffusion models; Layout [24], which manipulates the cross-attention layer in diffusion models to achieve precise spatial layout control; Structured [25], which enhances composability and attribute binding in text-to-image tasks by integrating linguistic structures with cross-attention layers; Attend-and-Excite [13], which improves semantic fidelity in text-to-image generation; and TokenCompose [15], which enhances text-to-image models through token-level supervision. The experimental results are presented in Table I, where MG, C, and F represent the MULTIGEN metrics, the COCO instance validation set, and the Flickr30k instance validation set, respectively.

The experimental results show that the baseline method, TokenCompose, achieves an average improvement of 12.8% over the Attend-and-Excite model across all the improved MG metrics. Since the FID metrics are reliable only when comparing 10,000 images, a comparison of the FID metrics on a dataset of 10,000 image-caption pairs sampled from the COCO validation set (C) shows a 3.3% reduction in the metrics of TokenCompose compared to the Attend-and-Excite model. The illustrative analysis results from the TokenCompose model further highlight the 12.8% improvement in MG metrics and the 3.3% reduction in FID metrics, representing substantial advancements in performance.

As shown in Table I, the proposed TI-OEAM model significantly outperforms all baseline methods across most evaluation metrics for these datasets. Compared to the state-of-the-art performance of current mainstream model, TokenCompose, on the COCO instance dataset, TI-OEAM achieves an average improvement of 12.36% on the MG2, MG3, MG4, and MG5 metrics. On the ADE20K instance dataset, TI-OEAM improves by an average of 6.55% on the same metrics. Additionally, TI-OEAM demonstrates a 0.84% reduction in the FID score compared to the 10,000 image-caption pairs in the COCO instance validation set. This performance enhancement is attributed to the model's efficacy, particularly its ability to learn information across multiple object categories through difficult sample construction and dynamic residual gating methods. Furthermore, TI-OEAM incorporates similarity constraints between the cross-attention map and the attention-weighted mask feature map, further improving the consistency between text prompts and the content of the generated images, resulting in significant gains across all performance metrics.

In comparison to the 1,000 image-caption pairs from the Flickr30K instance validation set, the FID metric for the images generated by the TI-OEAM model, as shown in Table I, is 71.94. The discrepancy in these experimental results can be attributed to the model's failure to pass the safety-checker detection during image generation based on the 1,000 captions. This issue led to the generation of a black image, which significantly affected the experimental outcomes.

E. Ablation Study

To validate the impact of the proposed OE and AN components on model performance, this paper conducts MG metrics ablation experiments on the COCO instance dataset and the ADE20K instance dataset, as well as FID metrics ablation experiments on 10,000 image-caption pairs sampled from the COCO instance validation set. Table II presents the experimental results, while Fig. 2 provides effectiveness analysis.

TABLE II. COMPARISON OF ABLATION EXPERIMENTAL RESULTS

Component	COCO INSTANCES				ADE20K INSTANCES			
	MG3	MG4	MG5		MG3	MG4	MG5	FID（C）
TI-OEAM	$79.43_{0.91}$	$32.55_{0.96}$	$4.32_{0.36}$		$79.39_{0.85}$	$37.13_{2.39}$	$7.04_{0.51}$	20.02
OE	$77.62_{1.08}$	$31.62_{1.26}$	$3.93_{0.46}$		$77.54_{0.77}$	$35.82_{1.30}$	$6.44_{0.63}$	19.90
AM	$78.48_{0.96}$	$31.44_{1.62}$	$3.91_{0.80}$		$79.18_{0.79}$	$37.80_{0.77}$	$7.62_{0.84}$	20.25

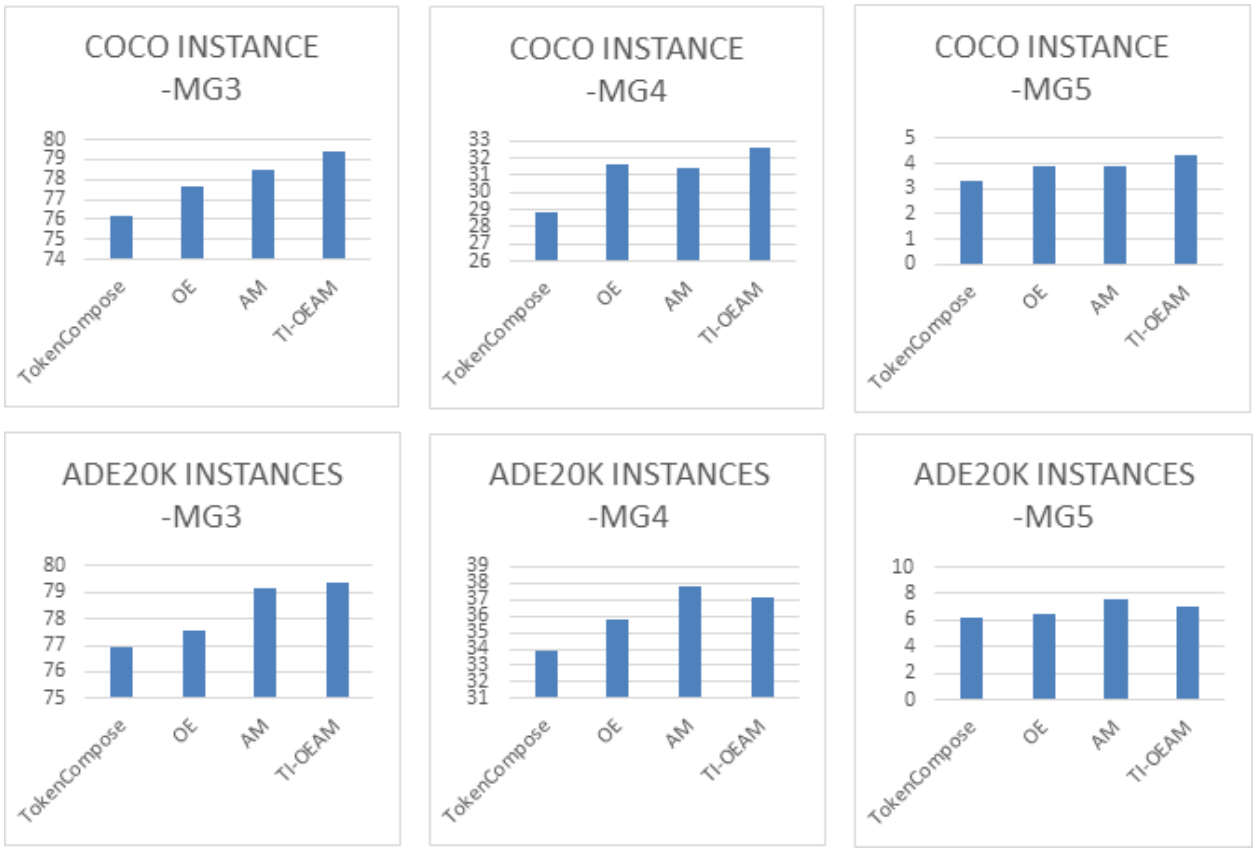

Fig. 2. Effectiveness analysis OE and AM.

Firstly, the OE module was independently validated, and the experimental results are presented in the "OE" section of Table II. Compared to the TokenCompose model, OE achieved improvements of 6.93% in the MG3, MG4, and MG5 metrics, respectively. This enhancement demonstrates that the OE module has made significant progress in extracting object features, thereby further enhancing the model's ability to learn object characteristics within the image.

Secondly, the AM method is validated in isolation, with the results presented in the "AM" section of Table II. Compared to the TokenCompose model, the AM method yields improvements of 11.41% in the MG3, MG4, and MG5 metrics. The AM method demonstrates significant improvements across all metrics, indicating that it enhances the model's understanding of the distribution of objects in images in alignment with text prompts. Furthermore, this method improves the consistency between text and image content, resulting in images with greater object accuracy.

In conclusion, the OE and AM modules demonstrate significant improvements in the metrics associated with the text-to-image generation task.

F. Visual Presentation Analysis

To facilitate a comprehensive visual comparison between the proposed TI-OEAM model and the benchmark TokenCompose model, a set of six images was generated using identical text prompts by both models. The generated images are presented in Fig. 3 on the following page for direct comparison. In order to ensure the fairness and consistency of the comparison, the initial latent space values, which serve as the starting point for both models, were held constant across all experiments. This approach minimizes potential bias arising from variations in latent representations, thereby allowing for an objective assessment of the models' performance.

Fig. 3. Qualitative comparison between TI-OEAM and baseline.

To rigorously evaluate the quality and alignment of the generated images with the text prompts, this study involved a manual assessment conducted by 11 researchers from diverse fields, including natural language processing, big data, and computer science. The researchers were tasked with evaluating the consistency between the provided text prompts and the corresponding generated images, using a well-defined criterion to ensure objectivity. The results of the evaluation revealed that, out of the total 66 votes cast, 26 votes were in favor of the images generated by the TokenCompose model, while 40 votes were in favor of those generated by the TI-OEAM model. This indicates that the images produced by the TI-OEAM model demonstrated a significantly higher degree of consistency with the text prompts compared to those generated by TokenCompose, highlighting the effectiveness of the proposed model in faithfully translating textual descriptions into visual representations.

IV. Conclusion and Future Work

This paper proposes a text-to-image generation method based on object enhancement and attention maps. The generation of high-quality images is achieved through the construction of challenging samples, the implementation of a dynamic residual gating mechanism, and the optimization of the model via an attention map guidance approach. These strategies work together to enhance the consistency between text prompts and generated images. Experimental results demonstrate that the proposed method, which integrates difficult sample design, dynamic residual gating, and attention map optimization, yields more significant improvements than state-of-the-art models in both MG metrics and the consistency of text-image information for text-image generation tasks. Future work will continue to address the consistency issue between text prompts and image contents, with a particular focus on more complex text prompts. The aim is to provide practical solutions for this challenge in the field.

Acknowledgment

This work is partially supported by Guangxi Innovation Driven Development Project (AA20302001).

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] Feng Z, Zhang Z, Yu X, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10135-10145.

[3] Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with cross attention control[J]. arXiv preprint arXiv:2208.01626, 2022.

[4] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//International conference on machine learning. Pmlr, 2021: 8821-8831.

[5] Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. 2023: 22500-22510.

[6] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in neural information processing systems, 2022, 35: 36479-36494.

[7] Xue Z, Song G, Guo Q, et al. Raphael: Text-to-image generation via large mixture of diffusion paths[J]. Advances in Neural Information Processing Systems, 2024, 36.

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

[9] Chen J, Yu J, Ge C, et al. Pixart-$\alpha $: Fast training of diffusion transformer for photorealistic text-to-image synthesis[J]. arxiv preprint arxiv:2310.00426, 2023.

[10] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[11] Liu N, Li S, Du Y, et al. Compositional visual generation with composable diffusion models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 423-439.

[12] Liew J H, Yan H, Zhou D, et al. Magicmix: Semantic mixing with diffusion models[EB/OL]. (2022-10-28)[2024-09-25].https://arxiv.org/pdf/2210.16056v1.pdf.

[13] Chefer H, Alaluf Y, Vinker Y, et al. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-10.

[14] Ma W D K, Lahiri A, Lewis J P, et al. Directed diffusion: Direct control of object placement through attention guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(5): 4098-4106.

[15] Wang Z, Sha Z, Ding Z, et al. Tokencompose: Grounding diffusion with token-level supervision[EB/OL]. (2023-12-06)[2024-09-25].https://arxiv.org/pdf/2312.03626v2.pdf.

[16] Lingenberg T, Reuter M, Sudhakaran G, et al. DIAGen: Diverse Image Augmentation with Generative Models for Few-Shot Learning[EB/OL].(2024-08-26)[2024-09-25].https://arxiv.org/pdf/2408.14584v1.pdf.

[17] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[18] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 633-641.

[19] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ade20k dataset[J]. International Journal of Computer Vision, 2019, 127: 302-321.

[20] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2641-2649.

[21] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[22] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[23] Loshchilov I. Decoupled weight decay regularization[EB/OL].(2019-01-04)[2024-09-25].https://arxiv.org/pdf/1711.05101v3.pdf.

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

[25] Feng W, He X, Fu T J, et al. Training-free structured diffusion guidance for compositional text-to-image synthesis[EB/OL].(2023-02-28)[2024-09-25].https://arxiv.org/pdf/2212.05032v3.pdf.