Multimodal Fusion For Text-to-Image Synthesis: A GAN Framework Driven By CLIP And CAM

Qiuyong Huang and Ailong Tang*

College of Information Science and Engineering, Liuzhou Institute of Technology, Liuzhou 545616, Guangxi, China

Corresponding author. E-mail: 110hqy@163.com

Received: Oct. 25, 2025; Accepted: May 08, 2026

Targeting the issues of weak fine-grained alignment capability and insufficient semantic controllability in existing generative adversarial network approaches, this paper presents a multimodal fusion-based model, namely Contrastive Language–Image Pretraining-Cross-Attention-Generative Adversarial Networks (CLIP-CA-GAN). With GAN as the basic architecture, this model incorporates the Contrastive Language-Image Pretraining (CLIP) model to establish multimodal semantic constraints. It dynamically fuses the local features of text and images via the Cross-Attention Mechanism (CAM), and optimizes generation quality through a designed Feature Fusion Module and a comprehensive loss function (LF). Experimental results demonstrate that the performance of CLIP-CA-GAN outperforms mainstream methods. On MS-COCO, the Fréchet Inception Distance (FID) decreases to 16.09, and the Inception Score (IS) rises to 4.91. On CUB, the FID stands at 14.06, the IS at 5.33, and the R–precision (RP) reaches 79.24. Additionally, the model has a relatively small number of parameters and high training efficiency, thus providing a high-quality and low-complexity solution for image generation.

Keywords: CLIP-CA-GAN, multimodal, CLIP, fine-grained alignment, CAM

http://dx.doi.org/10.6180/jase.202609_32.052

1. Introduction

Text-to-Image Generation (T2I) constitutes one of the central tasks within the realm of cross-modal generation [1]. Its objective is to produce semantically congruent images grounded in natural-language descriptions, facilitating profound collaboration between natural language processing and computer vision [2, 3]. T2I has found extensive application across diverse fields, including scene modeling [4], artistic creation [5, 6], intelligent mapping [7, 8], video generation [9, 10], and image analysis [11–13].

At present, T2I predominantly hinges on diffusion models, Transformers, or GAN, and its underlying rationale can be encapsulated within the "feature extraction-conditional injection" paradigm [14]. In other words, features are extracted via encoders and subsequently introduced into the generative network through means such as concatenation, addition, or attention mechanisms [15]. Nevertheless, the shortcoming of this paradigm lies in the asynchronous decoupling between the text semantic parsing and the image feature generation processes. Specifically, text features remain fixed and cannot be adjusted dynamically, and the generative network is unable to interpret text details adaptively. Moreover, the model lacks pre-trained prior knowledge and has to re-learn semantic-visual associations, which readily gives rise to semantic drift or attribute confusion.

More precisely, diffusion models generate images by incrementally adding noise and then reversing the process through denoising. Although they excel in overall photorealism and complex scene synthesis, and are stable during training [16], their implicit control over fine-grained seman-

tic alignment is tenuous. Dynamically binding specific text attributes to local image regions proves to be a challenging task. Even CLIP-guided diffusion models have yet to fully surmount the efficiency bottleneck and precise regulation issues [17, 18].

Conversely, GAN-based approaches are lightweight and efficient, yet their performance is also circumscribed by the aforementioned paradigm limitations. For instance, AttnGAN succumbs to semantic drift on account of its inability to dynamically adjust local details [19]; DM-GAN is prone to semantic confusion when generating multi-category objects [20]; DF-GAN struggles to capture attribute-level fine-grained associations [21]; ControlGAN has insufficient hierarchical semantic parsing for long texts [22]; and SSA-GAN's zero-shot generalization ability is restricted due to its dependence on additional annotated data [23].

In general, within the traditional "conditional injection" paradigm, text and image only interact at a superficial level or the input stage, lacking a prior knowledge-based dynamic semantic negotiation mechanism that pervades the entire generation process.

To address this issue, this paper proposes a prior-guided context-aware fine-grained generation method. The core of this method is to deeply integrate CLIP—originally an external feature extractor—into the semantic foundation of the generative architecture. This enables a shift from one-way conditional injection to two-way dynamic semantic negotiation, effectively enhancing fine-grained alignment capability. The core contributions of this paper are as follows:

1. A Fine-grained Alignment Generation Framework: With CLIP semantics as the context, CAM is employed to achieve dynamic word-region alignment, forming a progressive generation mechanism from global semantic constraints to local dynamic alignment. This effectively addresses the coarse-grained alignment problem of traditional models.

2. A Cross-modal Feature Fusion Mechanism: CAM is tasked with dynamically calculating the correlation weights between text words and image regions. These weights then guide the FFM to perform multi-level feature selection and enhancement, thus realizing end-to-end fine-grained control from semantic information guidance to feature fusion.

3. An Efficient and Lightweight Model Structure: While retaining the lightweight advantage of GANs, the model does not incur a significant increase in complexity after the introduction of semantic priors. Experiments show that CLIP-CA-GAN achieves a 12.7% increase in IS and an 18.3% decrease in FID on the CUB dataset. It has a relatively small number of parameters, with a training time of only 93.25 seconds per epoch.

1.1. Related Work and the Innovation Points

Table 1 compares CLIP-CA-GAN with related representative studies in terms of multi-modal alignment methods, feature fusion mechanisms, core innovations, and typical models.

The generation mechanism of CLIP-CA-GAN differs from that of mainstream models in the following aspects:

Interaction Depth: Most comparative models adopt shallow static interaction (AttnGAN) or external dynamic correction (Diffusion Models). In contrast, our model takes CLIP semantics as the foundation throughout the generation process and achieves continuous dynamic alignment with adaptive resolution adjustment via CAM.

Prior Knowledge: Comparative models use CLIP as an external feature extractor. However, our model integrates CLIP as an internal semantic context, enabling the model to acquire the awareness of "what should appear where".

Alignment Method: Comparative models adopt implicit data-driven alignment (GANs) or coarse-grained CLIP guidance (Diffusion Models). Our model calculates the word-region similarity via CAM and supplements it with FFM feature enhancement to achieve precise fine-grained attribute binding.

2. Methods

2.1. Model Structure

Fig. 1 depicts the architecture of CLIP-CA-GAN. The model consists of a generative network and a discriminative network. The generative network leverages CLIP to achieve fine-grained text-image alignment and embeds the CAM into key layers to integrate text semantics with image features. The discriminative network combines multi-head self-attention and multi-task loss to jointly optimize generation quality. The model takes images and text descriptions as inputs, outputs the authenticity probability and semantic matching score, updates parameters via backpropagation, and ultimately generates high-fidelity, semantically aligned images.

2.2. Contrastive Language–Image Pretraining

The multimodal semantic encoder adopts the pre-trained CLIP model. Through contrastive learning on a large corpus of text-image pairs, CLIP maps matched text and image data to an embedding space with consistent semantic fea-

Table 1. Comparative Analysis between the Model and Related Work.

Models	Generation Paradigm	Comparative Differences
AttnGAN, DMGAN	Data-driven conditional injection: Region-wise matching via embedded attention in GAN.	No prior guidance; attention learned from scratch; static interaction, weak generalization.
Diffusion+CLIP Guidance	External guidance: CLIP computes gradients to guide diffusion denoising.	CLIP external to generation network; high computational cost, coarse-grained alignment.
CLIP-GAN	Shallow conditional injection: CLIP textual features as GAN input.	Global alignment only; shallow text-image interaction, no fine-grained binding.
Masked Cross-Attention GAN [24]	Controllable focus: Mask-guided generation for specific regions.	Emphasis on controllable generation; lacks global-local consistency for generic descriptions; no CLIP prior context.
Diffusion Model (e.g., DALL•E 2, Imagen)	Iterative denoising generation	High image quality and rich details, but massive parameters and inference cost; coarse-grained text-image alignment due to implicit conditioning.
Transformer-based Models (e.g., Parti, Muse)	Autoregressive or masked prediction	Strong global consistency and scalability, yet lack explicit fine-grained word-region binding; typically require large-scale pre-training.
CLIP-CA-GAN	Prior-embedded dynamic negotiation: CLIP as semantic substrate, CAM for dynamic alignment, FFM for contextual fusion.	Method Innovation: Prior guidance + dynamic Negotiation:Mechanism innovation: CAM+FFM synergy;Efficiency Innovation: Lightweight and efficient.

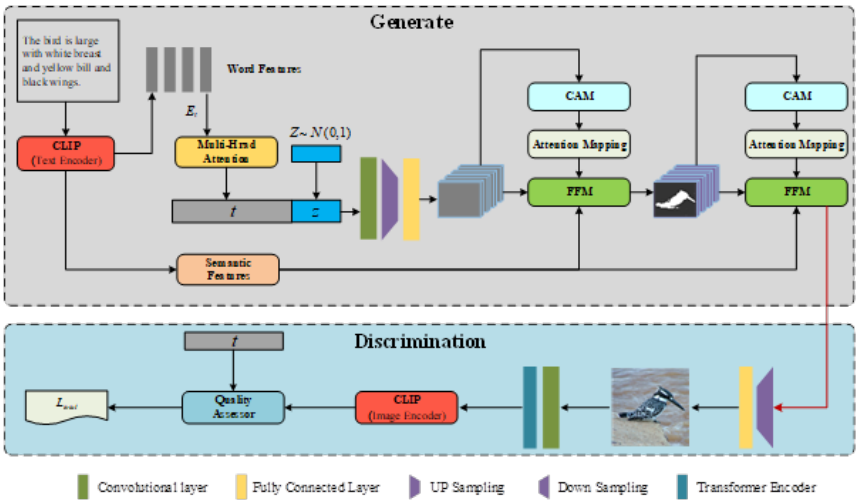

Fig. 1. Structure of CLIP-CA-GAN.

tures [25]. Fig. 2 shows the structure of CLIP, which mainly has two types of key outputs.

Global Semantic Embeddings: The text encoder transforms the input description T into a global semantic vector e_t, and the image encoder converts the image into a global vector E_{img} . These two are utilized to compute the semantic consistency loss (refer to Section 2.5).

Local Feature Sequences: The intermediate-layer features output by the image encoderspecifically, the feature sequence of image patches-provide fine-grained visual context for the subsequent CAM module.

2.3. Cross-Attention Mechanism

The aim of incorporating the CAM into the model is to deeply integrate words and image regions via dynamic interaction. This compensates for the shortcomings of CLIP in local alignment and generation controllability, thereby achieving complementary optimization of multimodal fea-

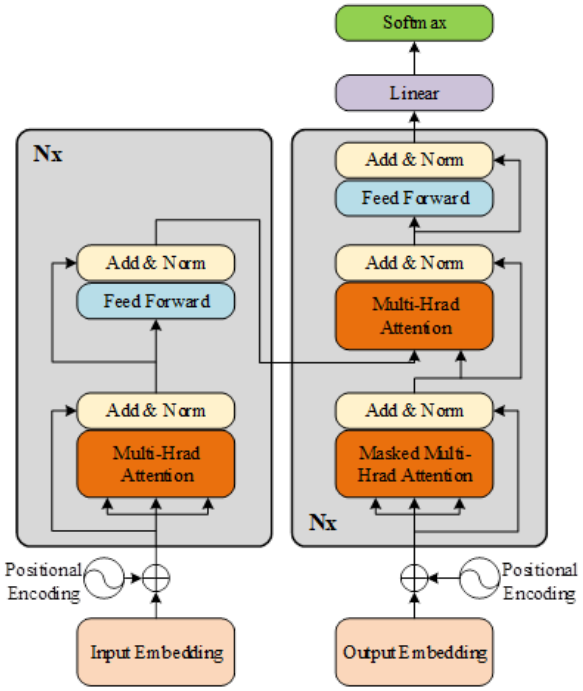

Fig. 2. Schematic diagram of CLIP structure.

tures [26].

First, word features are transformed into a common semantic space for image features, denoted as S, and the inner product is calculated between image features h and word features q, resulting in S′. Subsequently, the context vectors ci for each word in every sub-region of the image are computed based on the image's hidden features h′. The relevant calculation formulas are as follows:

$$S = f_{\text{text}}(q) \tag{1}$$

$$S' = h \cdot q \tag{2}$$

$$c_i = f_{\text{context}}(h', q_i) \tag{3}$$

In the equation, f_{text} represents the function that maps text features to the image feature space, q_i is the feature vector of the i-th word, and f_{context} is the function that calculates the word context vectors.

In the ensuing processing, the image features and the corresponding attention maps are input into the FFM. This module amalgamates multi-channel features, unifies their semantics and spatial distributions, and outputs optimized image features for utilization in the subsequent stage.

2.4. Feature Fusion Module

The structure of the FFM is illustrated in Fig. 3. It is composed of two sub-modules: the Feature Preprocessing Module and the Confidence Prediction Module.

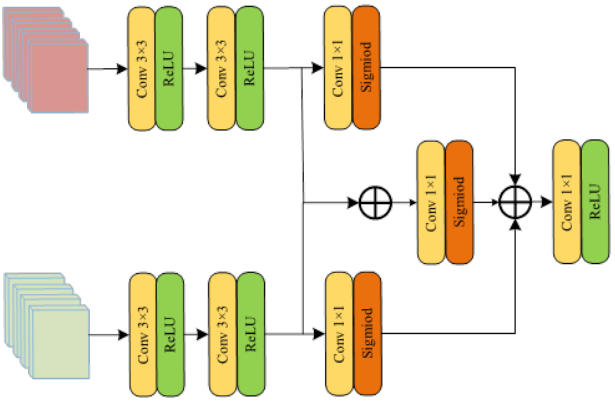

Fig. 3. Construction of FFM.

The Feature Preprocessing Module features a dual-branch structure, where each branch comprises two 3×3 convolutional layers and ReLU activation functions. This structure refines text semantic information and image features separately, thereby enhancing the feature representation capability of the network. A residual connection is additionally implemented between the input features and the output of the convolutional layers to preserve the original feature information and boost the network's learning capacity.

The Confidence Prediction Module has a triple-branch structure. Each branch contains a 1×1 convolutional layers and a Sigmoid activation function, which are utilized to compute the confidence scores of the input features and the fused features. Subsequently, the processed features are concatenated with the confidence values to generate a unified feature representation. The fused features incorporate the original semantics, attention information, and feature reliability assessment, offering richer and more precise feature support for subsequent image generation.

2.5. Total Loss

In addition to the adversarial loss, this paper utilizes cosine similarity as the semantic consistency loss to quantify the semantic consistency between the textual description and the generated image [27]. Furthermore, perceptual loss and L_1 loss are utilized as metrics to assess the quality of the produced images.

Semantic Consistency Loss:

$$L_{\text{sem}} = 1 - \frac{E_{\text{text}}(t) \cdot E_{\text{image}}(G(z,t))}{\|E_{\text{text}}(t)\|_2 \cdot \|E_{\text{image}}(G(z,t))\|_2} \tag{4}$$

Perceptual Loss:

$$L_{\text{perc}} = E_{x \sim P_{\text{real}}, z \sim P_z, t \sim P_{\text{text}}} [\|\phi(x) - \phi(G(z,t))\|_1] \tag{5}$$

L_1 Loss :

$$L_1 = E_{x\sim P_{\text{real}},z\sim P_z,t\sim P_{\text{text}}}\left[\|x - G(z,t)\|_1\right] \quad (6)$$

In the equation, $E_{\text{text}}(t)$ represents the semantic encoding of the text description; $E_{\text{image}}(G(z,t)$ denotes the semantic encoding of the generated image; x is the real image, following the real data distribution P_{real} ; z is the random noise vector, conforming to the noise distribution P_z; t is the text description, adhering to the text distribution P_{text} ; and $\phi(x)$ represents the feature extraction of the real image by the model.

Total LF is a weighted sum of the LFs mentioned above.

$$L_{\text{total}} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{perc}} L_{\text{perc}} + \lambda_{L1} L_1 \quad (7)$$

Here, λ is the weight coefficient of each loss function.

2.6. Pseudo Code

To better understand the working principle of CLIP-CA-GAN, the following provides its pseudocode, which details the main components and training process of the model.

3. Results and discussion

3.1. Dataset

Experiments were conducted on the CUB_200_2011 and MS-COCO datasets to train and evaluate the text-to-image generation model. The CUB_200_2011 dataset focuses on avian images and contains fine-grained categorical and attribute information. Its training set comprises 8,855 images and the test set 2,933 images, with ten text descriptions associated with each image [28]. The MS-COCO dataset encompasses a diverse range of daily scenes and objects [29]. with its training set including 82,783 images and the test set 40,504 images. Each image is provided with five text descriptions, and one text description was randomly selected as the target prompt during model training.

3.2. Experimental Setup and Evaluation Criteria

In terms of hardware, the GPU is an Nvidia RTX 409024 GB, which supports mixedprecision computing using both FP16 and FP32. The CPU is an Intel Xeon Platinum 8380H, including Intel AMX and AVX-512. The memory capacity is 128 GB .

The software environment is on the basis of the PyTorch 2.0.1 deep learning outline, coupled with CUDA 11.7 and cuDNN 8.5.0, supporting dynamic computation graphs and automatic mixed precision (AMP). The multimodal processing library uses HuggingFace Transformers 4.28.1, specifically for loading the CLIP model and text encoding. The initial model parameters are illustrated in Table 3.

All models are trained for 200 epochs with standardized settings. Random seed 42 ensures reproducibility (controls initialization and data order). Results are test-set based; each experiment is repeated five times, reported as mean ± std. Metrics include FID, IS, RP [30, 31].

3.3. Ablation Experiment

To investigate the impact of each module on model performance, ablation experiments were conducted on the COCO dataset. A standard GAN was used as the baseline model, with each enhancement module incrementally incorporated into the framework. Eight sets of experiments were performed to validate the superiority of CLIP-CA-GAN, and the results are presented in Tables 4(a) and 4(b). In the tables, " $\surd$ " indicates the inclusion of a module, while " × " indicates its exclusion.

As evidenced in Tables 4(a) and 4(b), Approach E outperforms all alternative methods across all evaluation metrics. Compared with Method A , Method B achieves a 23.1% reduction in FID and a 16.1% increase in IS. Method C further improves on Method B, with a 4.5% FID reduction and a 3.7% IS increase. Method D continues this improvement trend, reducing FID by 4.1% and increasing IS by 2.5% relative to Method C. Finally, Method E (CLIP-CA-GAN) achieves the lowest FID of 16.09 and the highest IS of 4.91, demonstrating a significant performance advantage.

Fig. 4 visualizes CAM's cross-layer attention weights, illustrating the evolution from global semantics to fine-grained local attribute alignment.

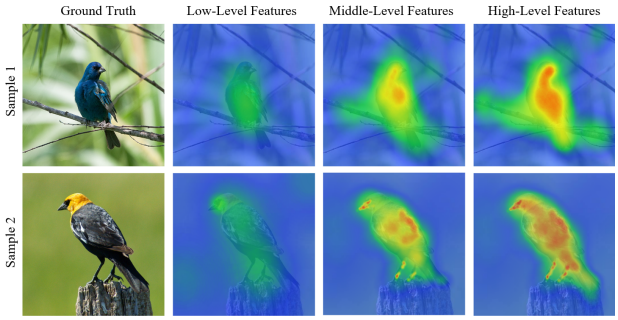

Fig. 4. Cross layer semantic alignment visualization.

For Sample 1, low-level attention maps cover the bird's contour; mid-level attention distinguishes between the body and head regions; high-level attention precisely focuses on the black legs and yellow neck/head, realizing color-to-part semantic binding.

For Sample 2, low-level attention covers the bird's body; mid-level attention differentiates between the wings and bill; high-level attention sharpens on black wing

Table 2. Comparative Analysis between the Model and Related Work.

Code 1: Training process of the CLIP-CA-GAN text-generated image model
Input: Random noise vector z, text description T
Output: Generated image in G_{img}
1. Text Encoding:
Extracting Text Semantic Features Using Pretrained CLIP Model; Text embedding $E_t = \text{CLIP}\ (\ T\ )$;
2. Generator forward propagation:
Input noise vector z and text embedding E_t;
Generate initial feature maps through multi-layer convolutional networks;
Calculate the correlation between text features and image features through the CAM, and output the fusion feature F_{fuse} ;
Generate image $G_{\text{img}} = \text{Generator\_Conv}\ (F_{\text{fused}}\ )$
3. Discriminator processing:
Input image G_{img} and text embedding E_t;
Extracting multi-scale image features through conv-layers;
Calculate the correlation weight between local regions and the global semantics of an image;
Output discrimination result: Authenticity probability D_{real} and semantic matching score $D_{\text{match:}}$
4. Use L_{total} to calculate the difference between the generated image and the real image.
5. Backpropagation and optimization:
Update the parameters $\nabla\theta_D\ (L_{\text{adv}}\ )$ and $\nabla\theta_G\ (L_{\text{total}}\ )$ of discriminator D and generator G;
Adaptive adjustment of learning rate using Adam optimizer;
6. Feature enhancement strategy:
Dynamically adjust the temperature parameter τ of the multimodal attention layer during the training process;
Perform random dropout on CLIP-embedded features to enhance robustness;
7. Output generated image:
Obtain the final image G_{img} through forward propagation of the generator.

Table 3. Model parameters.

Item	**Parameters**
Optimizer	AdamW. Learning rate $lr_G = 1 \times 10^{-4}$.
	Momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Weight decay $\lambda_{\text{wd}} = 0.01$
Training Strategy	Batch Size: Global batch size is 128 .
	Learning rate scheduling: Cosine Annealing, Cycle length $T_{\max} = 200$ epochs.
Regularization and Stability	Warmup: Linearly increase the learning rate to the initial value in the first 5 epochs.
	Gradient clipping: Global gradient norm threshold $\gamma = 0.1$.
	Dropout: Dropout rate in the middle layer of the generator is $p = 0.2$.
Data preprocessing	Text encoding: CLIP TE (ViT-B/32), maximum text length of 77 tokens, BPE segmentation.
	Convolutional layer: He is initialized normally.
Model initialization	Fully connected layer: Xavier normal initialization.
	Pre-training weights: The CLIP encoder loads OpenAI's official pre-training weights (ViT-B/32).

Table 4 ((a)). Experimental results of network module ablation (LF L_{adv}).

Method	CLIP	CAM	FFM	FID↓	IS ↑
A	×	×	×	28.17 ± 0.72	3.94 ± 0.15
B	√	×	×	21.65 ± 0.63	4.55 ± 0.04
C	√	√	×	20.67 ± 0.57	4.72 ± 0.05
D	√	√	√	19.83 ± 0.64	4.84 ± 0.08

Table 4 ((b)). Experimental results of LF ablation (based on structure D).

Method	L_{adv}	L_{sem}	L_{perc}	$L1$	FID ↓	IS ↑
D-1	✓	×	×	×	19.83 ± 0.31	4.84 ± 0.08
D-2	√	√	×	×	17.45 ± 0.24	4.88 ± 0.03
D-3	√	√	√	x	16.92 ± 0.23	4.89 ± 0.03
E	√	√	√	✓	16.09 ± 0.20	4.91 ± 0.03

bars/primaries and a short, pointed bill, reflecting the hierarchical parsing of complex textual descriptions.

3.4. Generate Image Quality Assessment

To evaluate the superiority of CLIP-CA-GAN, taking the CUB dataset as an example, the semantic consistency of

the model is evaluated by calculating the matching degree between the generated images and text descriptions. Table 5 shows the comparison of image generation quality of different models.

Table 5. Comparative Analysis between the Model and Related Work.

Model	FID↓	IS↑	R-precision↑
GAN-SC	24.65 ± 1.05	4.79 ± 0.27	-
ControlGAN	25.34 ± 1.12	4.58 ± 0.09	69.33 ± 2.15
MirrorGAN	54.97 ± 2.38	4.56 ± 0.05	57.67 ± 1.89
AttnGAN	23.98 ± 0.96	4.36 ± 0.03	67.82 ± 1.97
DT-GAN	16.35 ± 0.52	4.88 ± 0.03	-
DF-GAN	14.81 ± 0.41	5.10 ± 0.03	44.83 ± 1.62
DM-GAN	16.09 ± 0.48	4.75 ± 0.07	72.31 ± 2.03
RAT-GAN	13.91 ± 0.35	5.36 ± 0.20	42.73 ± 2.03
SSA-GAN	15.61 ± 0.43	5.17 ± 0.08	75.9 ± 1.58
ITSC-GAN	17.36 ± 0.55	4.63 ± 0.03	82.77 ± 2.21
CLIP-CA-GAN	14.06 ± 0.38	5.33 ± 0.05	79.24 ± 1.92

Upon analyzing the data in Table 5, it becomes evident that CLIP-CA-GAN exhibits distinct performance advantages. Its FID value is 14.06 ± 0.38, outperforming DF-GAN and DM-GAN, with a reduction of 5.06% and 12.62% respectively. This implies that by leveraging CAM to achieve semantic alignment within the CLIP embedding space, the mode-collapse problem is mitigated, and the generated images possess higher visual quality and distribution consistency, bearing a closer resemblance to real-world images.

The IS value of CLIP-CA-GAN is 5.33 ± 0.05. When compared to the second-best model RAT-GAN 5.36 ± 0.20, while maintaining a comparable IS value, the standard deviation is decreased by 60%, demonstrating that the model has a stable distribution-modeling capability. The R -precision value is 79.24 , representing a 4.4% increase compared to the baseline SSAGAN. Additionally, although this indicator is lower than that of ITSC-GAN, the IS of CLIP-CA-GAN has increased by 15.55%, indicating that while preserving semantic fidelity, the diversity of generated samples has not been compromised.

To further evaluate the generality of CLIP-CA-GAN, a referential comparison was conducted between it and mainstream diffusion models on the COCO dataset, and the results are presented in Table 6. It should be noted that the diffusion models were pre-trained with massive data, having billions of parameters, and were evaluated in a zero-shot manner. In contrast, CLIP-CA-GAN was trained from scratch only on the COCO dataset, with only 13.06 million parameters. Table 6 shows that the FID (7.27-11.84) of the diffusion models is superior to that of our model (16.09), but their training and inference costs are extremely high. In addition, the IS of CLIP-CA-GAN is 4.91 . While being lightweight and efficient, it achieves good generation quality, making it suitable for resource-constrained scenarios.

Table 6. Comparison of Image Generation Quality on COCO Dataset.

Model	Params	FID↓	IS↑
DALL•E 2	$\sim$ 6.5 B	10.39	-
Imagen	$\sim$ 3 B	7.27	-
Stable Diffusion	$\sim$ 1 B	11.2	-
CogView2	$\sim$ 4.5 B	13.5	-
CLIP-CA-GAN	13.06 M	16.09	4.91

Fig. 5 shows a comparison of the image generation effects of some models. The analysis results of the image generation quality of each model are as follows:

Feature Accuracy : Diffusion models (DALL•E 2, Imagen, Stable Diffusion) produce images with a high degree of realism and abundant details. However, their fine-grained semantic alignment capabilities are relatively weak, making it difficult to precisely map specific words in the text to relevant regions in the image. CLIP-CA-GAN, by leveraging CAM to achieve dynamic word-region alignment, can accurately generate the details emphasized in the text, with clear forms of tiny objects. In contrast, models like ATTN-GAN and ITSC-GAN have problems such as blurred or missing details.

Fig. 5. A Comparative analysis of images generated by various models.

Color Consistency : Diffusion models generate images with natural colors and smooth transitions. The model proposed in this paper, through the cooperation between the CLIP semantic foundation and CAM, enables more precise control of color distribution. The color regions are accurate, and the transitions are natural, thus avoiding

Table 7. Model performance comparison.

Model	Generator G		Discriminator D		Training time/epoch
	Params	FLOPs	Params	FLOPs	
MirrorGAN	27.51 M	1.043 G	29.77 M	1.174 G	317.09s
AttnGAN	16.42M	0.565 G	13.29M	0.406G	195.23s
DT-GAN	19.26M	0.749 G	14.62 M	0.582 G	253.95s
DF-GAN	19.45 M	0.681 G	22.48 M	0.905 G	303.82s
DM-GAN	14.83 M	0.431 G	19.17M	0.786G	221.36s
RAT-GAN	14.17 M	0.409 G	16.03 M	0.674 G	97.79s
SSA-GAN	16.94 M	0.606G	13.28 M	0.395 G	89.55s
ITSC-GAN	14.76M	0.489 G	16.55 M	0.652 G	94.25s
CLIP-CA-GAN	13.06M	0.388 G	15.21 M	0.464 G	93.25s

misalignment or overflow. Other comparative models have insufficient understanding of descriptions of mixed colors, resulting in rather harsh color transitions.

Text Relevance : The implicit alignment mechanism of diffusion models leads to a relatively weak correlation between text and image, with limited precision in semantic control. CLIP-CA-GAN, through explicit word-region alignment, generates images that are highly relevant to the text description and can accurately reflect details in complex descriptions. Other models have slightly weaker word-level feature parsing capabilities, presenting issues such as discrepancies between the image and the description, as well as local deformations.

3.5. Comparison of Model Complexity

To evaluate the model efficiency, Table 7 compares the complexity of each model in terms of Params, FLOPs and training time.

An analysis of Table 6 shows that the number of Params of the generator in CLIP-CAGAN is 13.06 M , which is substantially lower than that of models such as MirrorGAN, AttnGAN, and DF-GAN. When compared to the second-best model, RAT-GAN, it has decreased by 7.8%. The FLOPs of its generator is 0.388 G , which is 62.8% and 48.2% less than those of MirrorGAN and DT-GAN respectively.

In the discriminator, the number of parameters of CLIP-CA-GAN is 15.21 M , which is at a medium level. Although it is higher than that of SSA-GAN and AttnGAN, it is 48.9% lower than that of MirrorGAN. The FLOPs is 48.7% and 41.0% less than those of DF-GAN and DMGAN respectively.

The training time of CLIP-CA-GAN is 93.25 s/epoch, which is slightly higher that that of SSA-GAN (89.55 s) and ITSC-GAN (94.25 s). However, the difference is less than 4%, within the optimal range of the same order of magnitude. This indicates that through the collaborative optimization of CAM and CLIP feature fusion, the model achieves efficient utilization of computing resources while maintaining the generation quality.

3.6. Ethical considerations

T2I generation technology may potentially pose ethical risks in practical applications. The main risks and mitigation strategies are as follows:

Data Bias: Pre-trained models may have implicit biases, resulting in stereotypes in the generated results. Mitigation methods: Bias auditing, fine-tuning with balanced data, and informing users of uncertainties.

Content Misuse: It may be used to generate misleading or harmful information. Mitigation methods: Content filtering, adding invisible watermarks, complying with regulations, and explicitly prohibiting malicious uses.

Copyright Issues: The training data may involve copyrighted works, and the definition of copyright for the generated content is not yet clear.

4. Conclusions

In response to the bottleneck of fine-grained alignment in the text-to-image generation task, this paper proposes an innovative integration at the generation mechanism level and constructs the CLIP-CA-GAN model. Its core lies in the new mechanism of "dynamic semantic negotiation under prior guidance": By integrating CLIP as the semantic foundation and introducing CAM to perform continuous dynamic alignment in the CLIP space, the binding of text and image from global to local is achieved. Experimental results show that this model has excellent performance in terms of FID, IS, R-precision, and computational efficiency, verifying the effectiveness of the proposed solution in this paper. It not only endows GANs with powerful semantic understanding capabilities but also provides a lightweight dynamic alignment idea that can be borrowed by diffusion models and Transformers, which is expected to promote the further development of efficient and controllable cross-modal generative models.

It must be objectively noted that this study still has certain limitations. Firstly, the model's semantic understanding is restricted by the CLIP encoder, and it may perform suboptimally for rare concepts or complex combined semantics. Secondly, the current evaluation is centered around relatively standardized benchmark datasets, and the model's generalization ability in zero-shot scenarios involving highly complex scene compositions, diverse object interactions, and abstract visual descriptions requires further verification.

Future research directions are as follows: Firstly, explore a cascaded or hybrid architecture with diffusion models. Employ this model to generate semantic structures and combine them with the rendering capabilities of diffusion models for detail enhancement, aiming to strike a new balance between speed and quality. Secondly, consider the dynamic fine-grained alignment mechanism as a general conditional injection method and attempt to transfer and optimize the controllability of the Transformer-based autoregressive image generation model. Thirdly, extend the research from static images to temporal dynamic generation, and investigate the problem of maintaining cross-frame semantic consistency in text-to-video generation.

Competing interests

No competing interests are disclosed by the authors.

Authorship contribution statement

Ailong TANG: Supervision, Conceptualization, Project administration, Writing-Original draft preparation.

Qiuyong HUANG: Software, Methodology.

Data availability

Available upon request.

Declarations

Not applicable.

Conflicts of interest

The scholars state that they have no conflict of interest related to the publication of this paper.

Author statement

Every author has reviewed and approved the manuscript, confirming that the authorship requirements outlined earlier in this document have been met. Each author also believes that the manuscript accurately reflects their work.

Funding

This work is supported by 2025 Guangxi University Young and Middle-aged Teachers' Scientific Research Basic Ability Improvement Project Research on Multimodal Visual Fusion Based on Deep Learning (2025KY1156); by 2025 Special Project on Innovation and Entrepreneurship Education of Liuzhou Institute of Technology
AI Empowered Teaching Reform of Programming Courses—Exploration on Application-oriented Talent Training Through Integration of Industry and Education in Private Universities (2025SCZX06)

Ethical approval

All authors have been personally and actively engaged in significant efforts leading to the completion of this paper and will accept public responsibility for its content.

References

[1] F. Bie, Y. Yang, Z. Zhou, A. Ghanem, M. Zhang, Z. Yao, X. Wu, C. Holmes, P. Golnari, D. A. Clifton, et al., (2024) *"Renaissance: A survey into ai text-to-image generation in the era of large model"* **IEEE transactions on pattern analysis and machine intelligence 47**(3): 2212–2231. DOI: 10.1109/TPAMI.2024.3522305.

[2] V. Paananen, J. Oppenlaender, and A. Visuri, (2024) *"Using text-to-image generation for architectural design ideation"* **International Journal of Architectural Computing 22**(3): 458–474. DOI: 10.1177/14780771231222783.

[3] J. Oppenlaender. "The cultivated practices of text-to-image generation". In: *Humane Autonomous Technology: Re-thinking Experience with and in Intelligent Systems*. Springer, 2024, 325–349. DOI: 10.1007/978-3-031-66528-8_14.

[4] L. Höllein, A. Božič, N. Müller, D. Novotny, H.-Y. Tseng, C. Richardt, M. Zollhöfer, and M. Nießner. "Viewdiff: 3d-consistent image generation with text-to-image models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, 5043–5052. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Hollein_ViewDiff_3D-Consistent_Image_Generation_with_Text-to-Image_Models_CVPR_2024_paper.html.

[5] J. Gartner and M. Romanov, (2024) *"The advantages of ai text to image generation"* **International Journal of Art, Design, and Metaverse 2**(1): 1–8. DOI: https://topazart.info/e-journals/index.php/ijam/article/view/65.

[6] A. Dunkel, D. Burghardt, and M. Gugulica, (2024) *"Generative text-to-image diffusion for automated map production based on geosocial media data"* **KN-Journal of Cartography and Geographic Information *74***(1): 3–15. DOI: 10.1007/s42489-024-00159-9.

[7] A. A. Laghari, V. V. Estrela, and S. Yin, (2024) *"How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging"* **Current Medical Imaging *20***(1): e281222212228. DOI: 10.2174/1573405619666221228094228.

[8] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra, and M. Hoai. "Handiffuser: Text-to-image generation with realistic hand appearances". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 2468–2479. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Narasimhaswamy_HanDiffuser_Text-to-Image_Generation_With_Realistic_Hand_Appearances_CVPR_2024_paper.html.

[9] A. A. Laghari, Y. Sun, M. Alhussein, K. Aurangzeb, M. S. Anwar, and M. Rashid, (2023) *"Deep residual-dense network based on bidirectional recurrent neural network for atrial fibrillation detection"* **Scientific reports *13***(1): 15109. DOI: 10.1038/s41598-023-40343-x.

[10] A. A. Laghari, S. Shahid, R. Yadav, S. Karim, A. Khan, H. Li, and Y. Shoulin, (2023) *"The state of art and review on video streaming"* **Journal of High Speed Networks *29***(3): 211–236. DOI: 10.3233/JHS-222087.

[11] M. A. Munir, R. A. Shah, M. Ali, A. A. Laghari, A. Almadhor, and T. R. Gadekallu, (2024) *"Enhancing gene mutation prediction with sparse regularized autoencoders in lung cancer radiomics analysis"* **IEEE Access *13***: 7407–7425. DOI: 10.1109/ACCESS.2024.3523330.

[12] S. Karim, Y. Zhang, A. A. Laghari, and M. R. Asif. "Image processing based proposed drone for detecting and controlling street crimes". In: *2017 IEEE 17th International Conference on Communication Technology (ICCT)*. IEEE. 2017, 1725–1730. DOI: 10.1109/ICCT.2017.8359925.

[13] U. Saeed, K. Kumar, M. A. Khuhro, A. A. Laghari, A. A. Shaikh, and A. Rai, (2024) *"DeepLeukNet—A CNN based microscopy adaptation model for acute lymphoblastic leukemia classification"* **Multimedia Tools and Applications *83***(7): 21019–21043. DOI: 10.1007/s11042-023-16191-2.

[14] G. Marcus, E. Davis, and S. Aaronson, (2022) *"A very preliminary analysis of DALL-E 2"* **arXiv preprint arXiv:2204.13807**: DOI: 10.48550/arXiv.2204.13807.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 10684–10695. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html?utm_source=rns.dwaiai.de.

[16] H. Li and X.-J. Wu, (2024) *"CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach"* **Information Fusion *103***: 102147. DOI: 10.1016/j.inffus.2023.102147.

[17] R. Tamilkodi, K. Suryakala, N. Vamsi, M. Arvind Reddy, M. Nithin Kumar, and M. Venkat. "Transforming text into art: Exploring dall-e's text-to-image generation capabilities". In: *International Conference on Smart Data Intelligence*. Springer. 2024, 413–421. DOI: 10.1007/978-981-97-3191-6_31.

[18] Y. Shi, M. Shang, and Z. Qi, (2023) *"Intelligent layout generation based on deep generative models: A comprehensive survey"* **Information Fusion *100***: 101940. DOI: 10.1016/j.inffus.2023.101940.

[19] S. Naveen, M. S. R. Kiran, M. Indupriya, T. Manikanta, and P. Sudeep, (2021) *"Transformer models for enhancing AttnGAN based text to image generation"* **Image and Vision Computing *115***: 104284. DOI: 10.1016/j.imavis.2021.104284.

[20] L. Yan, R. Yan, B. Chai, G. Geng, P. Zhou, and J. Gao, (2024) *"DM-GAN: CNN hybrid vits for training GANs under limited data"* **Pattern Recognition *156***: 110810. DOI: 10.1016/j.patcog.2024.110810.

[21] R. Mehmood, R. Bashir, and K. J. Giri. "Comparative Analysis of AttnGAN, DF-GAN and SSA-GAN". In: *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE. 2021, 370–375. DOI: 10.1109/ICAC3N53548.2021.9725424.

[22] D. S. Patra and S. Padhee. "Comparative Analysis of ControlGAN and ControlGAN-GP Models based Text-to-Image Synthesis". In: *2022 OITS International Conference on Information Technology (OCIT)*. IEEE. 2022, 564–568. DOI: 10.1109/OCIT56763.2022.00110.

[23] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn. "Text to image generation with semantic-spatial aware gan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 18187–18196. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Liao_Text_to_Image_Generation_With_Semantic-Spatial_Aware_GAN_CVPR_2022_paper.html.

[24] S. Hou, Z. Li, K. Wu, Y. Zhao, and H. Li, (2024) *"Masked cross-attention and multi-head channel attention guiding single-stage generative adversarial networks for text-to-image generation"* **The Visual Computer *40***(12): 8639–8651. DOI: 10.1007/s00371-024-03260-2.

[25] S. A. Baumann, F. Krause, M. Neumayr, N. Stracke, M. Sevi, V. T. Hu, and B. Ommer. "Continuous, subject-specific attribute control in t2i models by identifying semantic directions". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, 13231–13241. DOI: https://openaccess.thecvf.com/content/CVPR2025/html/Baumann_Continuous_Subject-Specific_Attribute_Control_in_T2I_Models_by_Identifying_Semantic_CVPR_2025_paper.html.

[26] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, (2023) *"T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation"* **Advances in Neural Information Processing Systems *36***: 78723–78747. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/f8ad010cdd9143dbb0e9308c093aff24-Paper-Datasets_and_Benchmarks.pdf.

[27] T. Hu, L. Li, J. Van de Weijer, H. Gao, F. S. Khan, J. Yang, M.-M. Cheng, K. Wang, and Y. Wang, (2024) *"Token merging for training-free semantic binding in text-to-image synthesis"* **Advances in Neural Information Processing Systems *37***: 137646–137672. DOI: 10.52202/079017-4372.

[28] N. S. Mudiraj and S. Singh, (2025) *"Semantic mapping of Hindi text-to-image generation using CUB dataset"* **Scientific Reports *15***(1): 36632. DOI: 10.1038/s41598-025-20537-1.

[29] O. Durusoy et al., (2025) *"Open-source datasets for image processing and artificial intelligence research: A comparison of imagenet and ms coco datasets"* **Int. J. Sci. Innov. Eng *2***: 639–653. DOI: 10.70849/IJSCI0205202575.

[30] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, (2022) *"Cascaded diffusion models for high fidelity image generation"* **Journal of Machine Learning Research *23***(47): 1–33. URL: http://jmlr.org/papers/v23/21-0635.html.

[31] S. Ramzan, M. M. Iqbal, and T. Kalsum, (2022) *"Text-to-image generation using deep learning"* **Engineering Proceedings *20***(1): 16. DOI: 10.3390/engproc2022020016.