CIT. Journal of Computing and Information Technology, Vol. 33, No. 4, December 2025, 267–283
doi: 10.20532/cit.2025.1006253

Structurally Controllable Text-to-Image Generation for Architectural Images Using Structural Consistency Loss

YuanShuai Lan, Min Liao, Mo Chen and Yi Ou

School of Electronic Information Engineering, Geely University, Chengdu, Sichuan, China

Interior scene design is a comprehensive field involving space planning, color matching, furniture arrangement, texture expression and other aspects, aiming at designing an aesthetically pleasing and functional interior environment. The existing Vincentian graphical models often have problems such as chaotic spatial layout and disproportionate components when dealing with architectural images with strict structural requirements. In this study, we propose a structural consistency loss function to realize implicit structural control by constraining the spatial distribution and semantic alignment of the cross-attention graph of the Qwen-Image model. Specifically, it includes 1) designing the spatial concentration loss to induce the attention regions corresponding to architectural components to be more compact and focused, and 2) introducing the semantic alignment loss to reduce the similarity between the attention maps corresponding to different components, and to enhance the discriminative power of visual-semantic correspondence. This loss function is jointly optimized with the base loss to drive the model to spontaneously learn from the text and follow the underlying laws of the building structure. Experiments on the MMIS dataset show that the final model achieves an optimal performance of FID 11.06 and IS 34.82 and also performs best on the CLIPScore (0.869), a measure of graphic alignment. There is a significant improvement in the actual generation of building images in terms of scale coordination, rationality of component location and overall structural realism. Unlike the methods relying on external conditions, this method provides a scalable solution to achieve structure-aware image generation by relying only on textual cues, which promotes the practicalization of generative AI in the field of professional design, and provides effective technical ideas and methodological references for the in-depth application of text-generated image technology in the field of strong structural requirements such as architecture and design.

ACM CCS (2012) Classification: Computing methodologies → Artificial intelligence → Computer vision → Computer vision representations → Image representations

Applied computing → Arts and humanities → Architecture (buildings)

Keywords: text-to-image generation, architectural image generation, structurally controlled generation, attention mechanism, loss function design, attention mechanism

1. Introduction

Architectural image generation is an important task in the fields of architectural design, urban planning and cultural heritage preservation, and its goal is to generate architectural images with reasonable structure, accurate proportion and aesthetic value based on textual descriptions [1]. With the development of artificial intelligence technology, the diffusion models represented by Stable Diffusion, Midjourney, and DALL-E have demonstrated powerful capabilities in generalized text-to-image generation [2]. However, they still face core challenges when targeting specialized domains such as architecture with strict geometric and functional constraints. Images generated by existing models often suffer from structural instability, disproportion, and confusing spatial relationships [3], which are rooted in the inherent gap between the training goals of general-purpose models and the precise structural control required by

architecture, as well as the variety of training data, which makes it difficult for models to take into account any one proprietary domain. How to make the model learn the structural semantics in textual descriptions has become the current bottleneck in image generation in architecture.

To enhance the controllability of generation, researchers have proposed two main classes of methods. The first category is based on external conditional inputs, such as ControlNet [4] and T2I-Adapter [5] methods, which guide the generation by introducing additional control signals such as edge maps and depth maps. Although this type of methods can improve the spatial control accuracy, they rely on the production of high-quality conditional graphs, sacrificing the directness and flexibility of text-driven approach. The second category is based on internal attention optimization methods, such as Attend-and-Excite [6], which ensures the generation of specific semantic objects by adjusting the cross-attention graph. However, these approaches focus on "object existence" and lack explicit modeling of complex structural properties such as overall layout and component relationships.

In the task of architectural image generation, there are obvious deficiencies in the existing methods: (1) mismatch between the control method and the application scenario: architectural design emphasizes the direct transformation from textual concepts to the overall structure, and the frequent production of control charts is not in line with the workflow of the initial creative dispersion; (2) lack of constraints on the structure of the "reasonableness" of the existing methods. The existing methods are not embedded with the a priori knowledge of composition, proportion, perspective, *etc.* in architecture, which leads to the generation of results that "seem to be true" but "do not make sense". Therefore, exploring a new method that can embed structural constraints without external conditions and only through textual hints has important theoretical value and application prospects.

To address the above shortcomings, the core objective of this paper is to enable the generic Vincennes graphical model to acquire the intrinsic perception and generation capability of building structures by designing a new training constraint mechanism without relying on any external condition signals. The main innovations of this paper are as follows:

1. Proposal of a structural consistency loss function to implicitly encode structural generation rules by constraining the spatial concentration and semantic alignment of the cross-attention graph and guiding the model to learn a robust mapping between architectural components in text and spatial regions in images.
2. Explore a "text-only" generation method, *i.e.*, extract control signals only from the textual cues themselves, and realize the controlled generation without external inputs, which significantly improves the structural reasonableness of the architectural images while maintaining the freedom of generation.

The remainder of the paper is organized as follows. In Section 2, systematic review of the related work on Vincentian graphs and controlled generation methods is performed. Section 3 elaborates on the design motivation, mathematical definition and implementation details of the proposed structural consistency loss function. Section 4 describes the experimental setup, including dataset processing, baseline modeling, and evaluation metrics, and conducts a comprehensive analysis of the results and ablation experiments. Section 5 analyzes the results of the study, compares them with the previous studies and explains the comparison methodology. Section 6 summarizes the full work and discusses the limitations of the current methodology and future research directions.

2. Related Review

2.1. Text-to-Image Generation Model

Text-to-image generation models are based on diffusion techniques to generate images based on textual cues. Throughout this research, the initial AttnGAN [7] family of models was used to establish associations between text and image regions through the attention mechanism, however, it was limited in generative diversity due to unstable Generative Adversarial Network (GAN) training and crash problems. Subsequent potential diffusion models, represented by Sta-

ble Diffusion [8], strike a balance between generation quality and computational efficiency, laying the foundation for subsequent research. These models rely on the CLIP [9] text encoder to encode cue words as semantic vectors and achieve graphical semantic alignment through a cross-attention mechanism. Recently, large language models, represented by the Qwen family, have gained strong generalized knowledge representation capabilities by pre-training on massive data. Among them, visual-linguistic bigram models such as Qwen-Image [10] and LAION-5B [11] further extend this capability to the multimodal domain, which can process and understand image and text information simultaneously, and further fuse deep text understanding and image generation capabilities in a unified Transformer framework, showing strong generality. Although these models perform well in the general-purpose domain, when applied directly to specialized domains such as interior design, there is still a certain deficiency in the accuracy of the generated images when compared to specialized images.

2.2. Controllable Image Generation Technology

In order to achieve accurate and high-quality control of the generated images, existing research has been carried out in three main technical paths, each with its own advantages and limitations.

2.2.1. A Method that Relies on External Condition Inputs

When there is a posing requirement for the generated image, using this method allows for the introduction of a learnable external adaptation controller while keeping the generative capabilities of the Vincentian graph model intact. This controller is responsible for parsing additional user-supplied spatial conditions (*e.g.*, edge maps, depth maps, semantic segmentation maps) and encoding them into control signals aligned with the model's internal noise latent images, thus enabling pixel-level spatial constraints. ControlNet is the seminal work in this paradigm, which constitutes a bypass branch via a trainable copy with structural symmetry and weight-locked weights of the original U-Net encoder that achieves accurate injection of multiple conditions. Subsequent studies such as T2I-Adapter and ControlNeXt [12] have been devoted to improving control efficiency and flexibility, *e.g.*, PixArt-δ [13] migrated ControlNet ideas to the Transformer backbone and combined with the potential consistency model to reduce the number of sampling steps to 4, realizing the "seconds" of controlled generation. However, the performance of such methods is strictly dependent on the availability of high-quality external conditions. For creative tasks such as architectural design that require starting from free-text ideas, pre-preparing precise condition maps instead constitutes a workflow bottleneck, limiting the freedom of design exploration.

2.2.2. Methods to Optimize Internal Model Mechanisms

When it is not possible to make improvements in the generated images through external methods, it is a good approach to make modifications directly inside the model. This approach enables more flexible control by directly intervening in the forward reasoning process inside the model by modulating the cross-attention mechanism of text-image association. For example, Attend-and-Excite ensures that all subject tokens in the cue are fully activated in the attentional layer through optimization; Paint-with-Words [14] allows the user to specify the correspondence of text tokens to image regions for local editing; and GLIGEN [15] anchors linguistic descriptions to spatial features through gating mechanisms. These types of approaches maintain the smoothness of the authoring process by starting directly from textual cues. However, they lack explicit mechanisms to force the generated results to conform to the structural relationships, scale and geometric consistency necessary for specialized domains (*e.g.*, architecture), and suffer from manual manipulation, which may result in results that are visually plausible but untenable in a professional sense.

2.2.3. Internal Knowledge Editing and Adaptation Technology

When a generic model lacks knowledge of specific concepts (*e.g.*, personalized objects, unique styles), the method implants new conceptual representations for the model by making lightweight, localized edits to its internal weights or embedding space. Represented by LoRA [16], it enables the model to efficiently learn and stably generate new concepts while avoiding catastrophic forgetting by injecting a trainable low-rank decomposition matrix into the weight matrix of the Transformer attention module. Multiple LoRA modules can be linearly combined for multi-concept blending during inference. This technique enables a high degree of personalization, but the granularity of its control usually lies in the text itself rather than in its precise spatial structural relationships.

2.3. This Work

In summary, with the existing methods, it is difficult to ensure that the generated results satisfy the strong structural constraints of the specialized domain based on textual descriptions alone. Methods that rely on external conditions are highly accurate but not flexible enough; methods that optimize internal mechanisms are flexible but have weak structural constraints. Therefore, how to generate high-quality architectural images without relying on any additional conditional inputs and only through textual cues is a key issue.

The aim of this work is to study and explore this problem and propose a tractable solution. This study proposes a structural consistency loss function that is dynamically optimized during the sampling process of the diffusion model. Unlike external control methods, it does not require any auxiliary input; and unlike generic internal optimization methods, it implicitly injects a structural coherence a priori by explicitly steering the cross-attention graph, prompting the spatial concentration of textual markers describing the key architectural components, and driving the semantic separation between different components. This enables the adaptation of a powerful generic visual-linguistic model to architectural image generation tasks without relying on architectural drawings or modifying the core architecture of the model, providing a new path towards flexible and structure-aware textual graph generation.

3. Method

3.1. Overall Framework

When generating large models using generic text-to-image generation, the problem of generating architectural images by inputting textual prompts is often structurally illogical and is rooted in the model's lack of explicit constraints on the structural principles of architecture. In order to solve the accuracy of the model in building image generation, this paper proposes an improved method of building image generation based on Qwen-Image model and MMIS dataset [17]. The overall framework is shown in Figure 1, and the core innovation lies in the introduction of a structural consistency loss function, which does not change the original model structure by constraining the attention mechanism distribution within the model during the training process, thus reducing the cost of the model; instead, a regularization term of the attention mechanism based on the location of the building is introduced in the training to guide the model to learn the structural rules of each component in the building image.

The overall process is as follows: first, the text prompt is put into the Qwen-Image model to start the diffusion generation process; when the model is forward calculated, the text prompt is put into the Qwen-Image model to start the diffusion generation process, and the text prompt is put into the Qwen-Image model to start the diffusion generation process, the attention weights between text and image features are extracted synchronously from the cross-attention layer of DIT module. Based on these weights, the spatial concentration loss and semantic alignment loss are calculated, which together constitute the structural consistency loss. Finally, this loss is weighted with the original diffusion loss to obtain the total training loss. The model is continuously optimized to generate images with reasonable building structures.

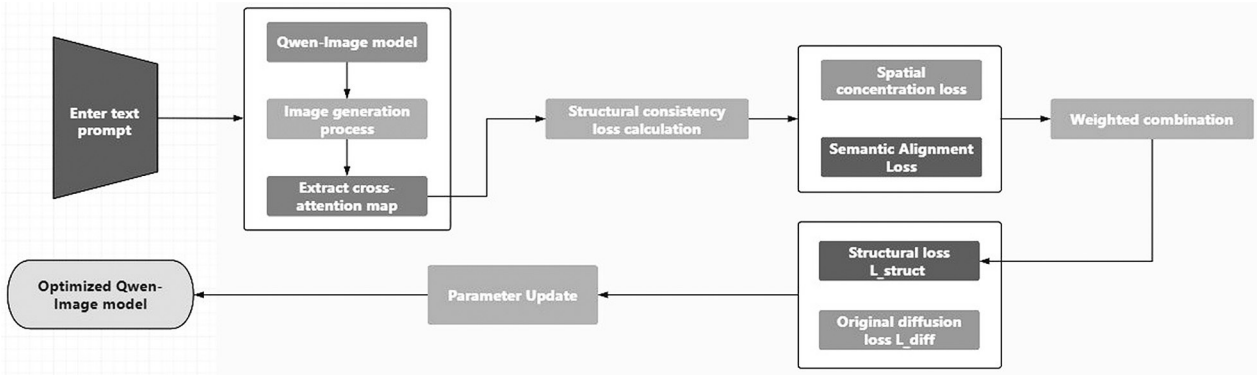

Figure 1. Qwen-Image Model with Loss Introduction.

Qwen-Image is a text-to-image generation model based on the Diffusion Transformer (DiT), and its overall architecture consists of three core components: the Text Encoder, the Diffusion Transformer (DiT), and the Variable Auto-Encoder (VAE). The text encoder adopts the Qwen2.5 visual language model (Qwen2.5_VL), which contains a 32-layer visual coder and a 28-layer language model to process image and text inputs, respectively. The diffusion transformer part contains 60 QwenImageTransformerBlocks, and the QwenDoubleStreamAttention mechanism is integrated within each block for multilevel fusion of text and image features. The VAE coder and decoder use a 3D convolutional architecture, which is responsible for coding and reconstruction of the image latent space. In this study, the attention graph extraction mechanism captures the attention weight tensor of shape (B, H, L, S) in real time by registering forward propagation hooks to the dual-stream attention module of layers 12 to 16 of the DiT module, which provides the basis for the computation of the structural coherence loss. The structure of Figure 1 is refined to obtain the detailed network structure as shown in Figure 2.

3.2. Attention Map Extraction Mechanism

The attention graph is the basis for constructing structural constraints. During the training process, attention weights are extracted in real time from the cross-attention layer of the DiT module via forward propagation hooks to obtain the strength of association between text tokens and image spatial locations.

Specifically, there are H attention heads in the cross-attention module in the l layer of the DiT model. With the acquired text features $X_l \in \mathbb{R}^{B\times L\times D}$ and image features $Y_l \in \mathbb{R}^{B\times S\times D}$, for the h-th attention head $h = 1, \ldots, H$, independent queries are first performed with key projection computation:

$$Q_l^h = X_l\, W_Q^{l,h},\ K_l^h = Y_l\, W_K^{l,h}. \tag{1}$$

$W_Q^{l,h}, W_K^{l,h} \in \mathbb{R}^{D\times D_h}$ are the learnable query and key projection matrices corresponding to the l-th level of the h-th attentional header, respectively, and D_h is the eigendimension of each header (usually $D_h = D/H$).

The weight matrix for this attention header is then computed based on the Query-Key:

$$A_l^h = \mathrm{softmax}\left(\frac{Q_l^h (K_l^h)^T}{\sqrt{D_h}}\right) \in \mathbb{R}^{B\times L\times S}. \tag{2}$$

The formula $A_l^h[b, i, j]$ denotes that in the b-th sample $b = 1, ..., B$, the i-th token of the text $i = 1, ..., L$ pairs with the j-th spatial bit of the image $j = 1, ..., S$ in the h-th attention weight of the head.

The outputs of all the attention heads are stacked in the second dimension, *i.e.*, the complete four-dimensional attention tensor of the l-th layer is obtained:

$$A_1 = \mathrm{stack}\left([A_l^1, A_l^2, ..., A_l^H], dim = 1\right) \in \mathbb{R}^{B\times H\times L\times S} \tag{3}$$

Its element $A_l\,[b, h, i, j]$ accurately represents the attention weight of the text's i-th token to the image's j-th spatial location in the b-th sample, h-th attention head.

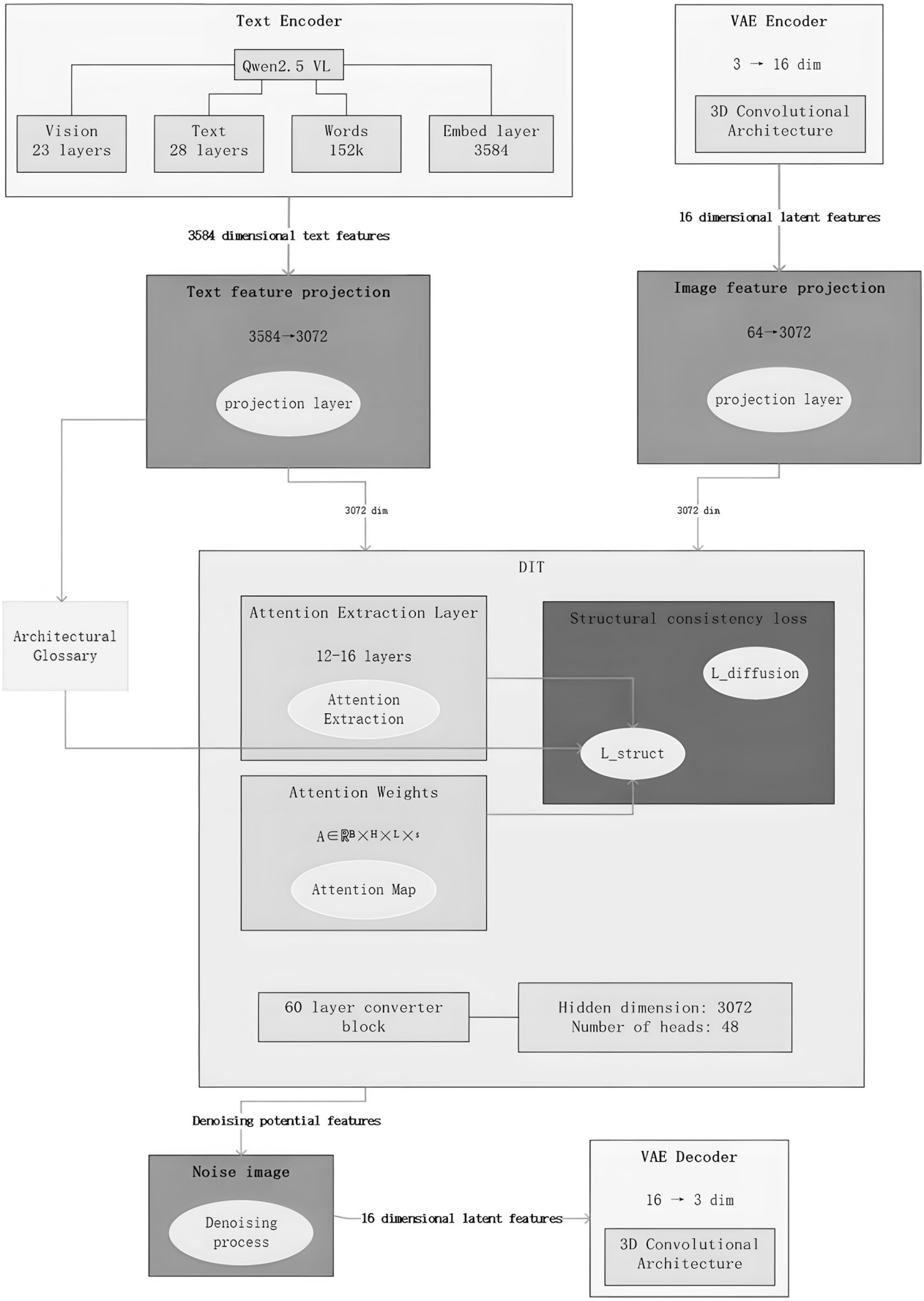

Figure 2. Detailed network structure.

By registering PyTorch forward propagation hooks into the QwenDoubleStreamAttention module of the Qwen-Image model, the fine-grained association between text and image spatial locations is obtained by capturing the attention tensor A_l in real time during the training process. It provides processing basis for the subsequent Spatial Concentration Loss and Semantic Alignment Loss.

3.3. Spatial Concentration Loss

In building image generation, specific building components (such as "Windows", "Doors", *etc.*) should correspond to specific local areas in the image, rather than diffuse distribution. In order to strengthen this positioning characteristic, the spatial concentration loss is designed to make the model focus on a specific local area rather than the whole image when dealing with building keywords.

That is, the attention distribution of a building component corresponding to a specific physical component should be spatially compact and continuous, with its probability mass concentrated in a coherent region. For the set of building component tokens K identified in the textual cue words, their cross-attention maps are extracted at the key layer of the denoising network to perform the following computational process:

1. Extraction and normalization: their attention vectors are obtained and reshaped into a two-dimensional matrix $M_t \in \mathbb{R}^{H\times W}$, and the spatial probability distribution maps are obtained by a Softmax function controlled by the temperature parameter τ:

$$P_t(i,j) = \frac{\exp(M_t(i,j)/\tau)}{\sum_{i',j'} \exp\left(M_t(i',j')/\tau\right)} \tag{4}$$

   where τ is used to adjust the smoothness of the distribution at the beginning of the optimization.

2. Calculation of spatial moments:

   The center of mass (first order moments) is given as:

$$\mu_x = \sum_{i,j} P_t(i,j)\cdot j,\ \ \mu_y = \sum_{i,j} P_t(i,j)\cdot i \tag{5}$$

   The variance (second-order central moments) is given as:

$$\sigma_x^2 = \sum_{i,j} P_t(i,j)\cdot (j-\mu_x)^2 \tag{6}$$

$$\sigma_y^2 = \sum_{i,j} P_t(i,j)\cdot (i-\mu_y)^2 \tag{7}$$

3. Defining loss: The loss of spatial concentration is defined as the mean of the variance of all target markers:

$$\mathcal{L}_{spatial} = \frac{1}{|K|}\sum_{t\in K}\left(\sigma_x^2+\sigma_y^2\right) \tag{8}$$

By minimizing $\mathcal{L}_{spatial}$, the model is explicitly steered during the training process so that the generated signals of the building components are constrained to be within a local region centered on their attentional centers of mass, which ensures higher spatial localization accuracy and structural integrity in the generation of building components. The set of building components is not combined with the word list of the original model but is obtained through textual cue words and attention extraction.

3.4. Semantic Alignment Loss

In the task of fine-grained text-to-image generation, ensuring that different semantic concepts are accurately mapped to different spatial regions in the image is a key challenge in achieving high-fidelity generation. In the case of architectural image generation, in order to distinguish the visual regions of specific components such as "window", "door", *etc.*, the cross-attention maps corresponding to each architectural component (Token) need to reduce the similarity between the attention maps corresponding to different architectural components. For each building component (Token), it is necessary to reduce the similarity between the attention maps corresponding to different building components, so that the model can assign more separate and exclusive attention regions to them in the image space during the generation process, thus enhancing the independence and localization accuracy of each component in the final image.

For a collection of building component token indices K identified in a batch, let $P_i \in \mathbb{R}^S$ denote the vector of attention graphs after flattening corresponding to the keyword i (with S being the number of spatial locations), and $|K|$ is the number of building components identified in the batch. The semantic alignment loss is defined as the average of the cosine similarity of the attention graphs between all different keyword pairs:

$$\mathcal{L}_{semantic} = \frac{2}{|K|(|K|-1)} \sum_{i \in K} \sum_{j \in K, j>i} \frac{\widetilde{P}_i \cdot \widetilde{P}_j}{\left\|\widetilde{P}_i\right\|_2 \cdot \left\|\widetilde{P}_j\right\|_2} \quad (9)$$

In order to eliminate the bias of different attention maps in the range and distribution of values, and to ensure the fairness and stability of similarity comparisons, we normalize each attention vector before computation:

$$\begin{aligned} \widetilde{P}_k &= \frac{P_k - \mu_k}{\sigma_k + \varepsilon}, \\ \mu_k &= \frac{1}{S} \sum_{S=1}^{S} P_k^{(S)}, \\ \sigma_k &= \sqrt{\frac{1}{S} \sum_{S=1}^{S} \left(P_k^{(S)} - \mu_k\right)^2}. \end{aligned} \quad (10)$$

where ε is a small constant used to maintain numerical stability. μ_k is the mean of the vector P_k, and σ_k is the standard deviation of the vector P_k.

3.5. Overall Structural Consistency Loss

To synthesize attention concentration and differentiation optimization, the above two loss terms are combined to obtain a total structural consistency loss:

$$L_{struct} = \alpha \cdot L_{textspatial} + \beta \cdot L_{semantic} \quad (11)$$

Where α and β are the balance hyperparameters; these two hyperparameters are used to jointly guide model learning to conform to architectural principles for structural representations.

3.6. The Total Loss Function

Adding structural consistency loss as a regular term to the original diffusion loss results in the total loss function:

$$L_{total} = L_{diffusion} + \lambda \cdot L_{struct} \cdot textstruct \quad (12)$$

Where $L_{diffusion}$ is the raw training loss (MSE loss) of Qwen-Image, λ controls the strength of the structural constraint. This combination ensures that the model maintains its original generation capabilities while specifically optimizing the structural rationality of building image generation through the guidance of structural consistency loss.

4. Results

4.1. Experimental Setup

4.1.1. The MMIS Dataset

In this experiment, we use the MMIS dataset, which is a large-scale dataset specially constructed for advancing multimodal indoor scene understanding and generation. The dataset is constructed with multimodal alignment data, containing nearly 160,000 high-quality indoor scene images, and each image is accurately paired with a detailed textual description and a voice recording of the description, forming an "image-text-audio" trinity data structure. The textual descriptions in this dataset accurately and meticulously reflect the visual elements and spatial relationships in the images, while the corresponding audio recordings further increase the modal diversity and potential research dimensions. The data consists of 11 decorative styles, namely Art Deco, Bohemian, Coastal, Contemporary, Eclectic, Farmhouse, Mediterranean, Mid-Century Modern, Rustic, and Traditional. In the experiment, the dataset is divided into a training set, a validation set and a test set according to the ratio of 7:2:1. The training set is used for learning and optimizing the model parameters, the validation set is responsible for monitoring the training process, and the test set is used for objectively assessing the generalization ability of the final model.

4.1.2. Model Training Configuration

The model was trained based on data shown in Table 1 and Table 2, and LoRA was used for style migration. The considered models were:

Table 1. Model training parameters.

Configuration Item	Configuration Content
Optimizer	AdamW
Batch Size	6
Epochs	50
Learning rate	1e-4
Hardware environment	10 × NVIDIA A100 GPU (100 GB)
Memory Optimization	Gradient checkpointing techniques, mixed precision training (BF16)

1. Stable Diffusion 3 (SD3): open-source, baseline model.
2. Qwen-Image (raw): native model not fine-tuned on architectural data.
3. Qwen-Image + FT: Base model fine-tuned on the MMIS dataset only.
4. Our Full Model: fine-tuned base with structural consistency loss for joint training.

4.1.3. Evaluation Metrics

In order to comprehensively assess the quality of the generated images, a multi-dimensional evaluation system was used in this study:

1. IS (Inception Score) [18]: assesses the quality and diversity of generated images by pre-training the Inception network, the higher the value the better.
2. FID (Fréchet Inception Distance) [19]: measures the distribution distance between the generated and real images in the feature space of the Inception network, with lower values indicating higher visual realism.
3. KID (Kernel Inception Distance) [20]: as an unbiased estimation of FID, also used to assess distributional similarity, more robust to sample size, lower values are better.

Table 2. LoRA Configuration Parameters.

Configuration Item	Configuration Content
Target module	to_q, to_k, to_v, add_q_proj, add_k_proj, add_v_proj, to_out.0, to_add_out, img_mlp.net.2, img_mod.1, txt_mlp.net.2, txt_mod.1
Lora rank	32
Lora alpha	64
Dropout	0.1

4. CLIPScore [21]: based on the cosine similarity between the generated image and the input text computed by the pre-trained CLIP visual-linguistic model, directly reflecting the degree of semantic alignment between the image and the input text, the higher the value the better.
5. LPIPS (Learned Perceptual Image Patch Similarity) [22]: calculates the similarity between image pairs based on perceptual features to assess detail fidelity and structural consistency, the lower the value the more similar perceptually, in this study VGG network is used to do the calculation.

4.2. Results Analysis

4.2.1. Comparative Experiment

The MMIS training set was used to train the model and the quantitative evaluation results obtained from the validation on the validation set using MMIS are shown in Table 3.

From Table 3, it can be seen that the fine-tuned model (Qwen-Image + FT) significantly outperforms both Qwen-Image and Stable Diffusion 3 in terms of FID and IS; whereas Our Full Model achieves the best performance with the lowest FID/KID (11.06/6.23) and the highest IS (34.82). It shows that the generated image is optimal in terms of visual realism, diversity and clarity; the highest CLIPScore (0.869) proves that Our Full Model is able to obtain better graphic semantic alignment in terms of textual cue word comprehension; and the lowest LPIPS (0.352) verifies that the proposed loss function effectively improves the structural reasonableness and compositional accuracy of the image by these indicators. Thus, it is possible that Our Full Model has a more accurate and reasonable image generation of buildings in a multi-dimensional way.

The change in loss during model training is shown in Figure 3. In the early stage of training, due to the introduction of the structural consistency loss function, the model needs to adapt to the new optimization objectives and constraints, and the loss value appears to rise briefly. With the increase in the number of iterations, the model gradually learns an effective representation of the building structure, and the loss starts to decrease steadily. Eventually, the loss curve converges and stabilizes, indicating that the model has sufficiently adapted to the new architecture and is able to stably generate structurally sound building images. This training dynamic confirms the optimizability of the proposed loss function and its benign guiding effect on the model convergence process.

4.2.2. Ablation Experiments

In order to verify the specific contribution of each loss component, ablation experiments were designed on the validation set to directly respond to the degree of semantic alignment between the image and the input text with the CLIPScore metric, and the distribution distance between the generated image and the real image with the FID metric. The results were obtained as shown in Table 4.

From the analysis in Table 4, it is evident that the addition of spatial concentration loss (Qwen-Image + $L_{spatial}$) alone significantly improves the spatial consistency of the building compo-

Table 3. Quantitative evaluation results of different models on the MMIS test set.

Model	FID	KID	IS	CLIPScore	LPIPS(VGG)
Stable Diffusion 3	18.73	12.45	25.41	0.792	0.421
Qwen-Image	15.82	14.82	24.18	0.815	0.438
Qwen-Image + FT	13.45	8.91	31.25	0.841	0.385
Our Full Model	11.06	6.23	34.82	0.869	0.352

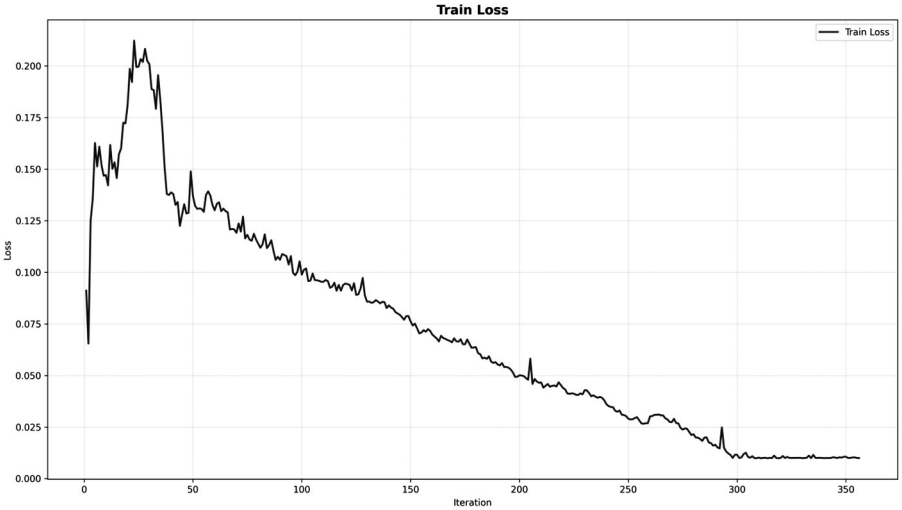

Figure 3. Model Training Losses.

Table 4. Results of ablation experiments.

Model Configuration	CLIPScore	FID
Qwen-Image + $L_{spatial}$	0.833	11.73
Qwen-Image + $L_{semantic}$	0.852	12.82
Our Full Model	0.869	11.06

nents, but has limited direct enhancement of the graphic alignment; the addition of semantic alignment loss (Qwen-Image + $L_{semantic}$) alone is most effective in improving the graphic alignment (CLIPScore). The combination of the two (Our Full Model) achieves complementarity and synergy, optimizing on all metrics and making the images generated from textual cue words spatially and semantically consistent with reality. This demonstrates that the two loss functions are explicitly complementary: the spatial loss dominates layout optimization and the semantic loss dominates association enhancement, and the two work synergistically to drive the model towards a more structurally and semantically aligned building image generation.

4.3. Qualitative Results

Figure 4 shows a comparison of the results of different models for the same textual cues. For example, for the prompt "Art Deco, a bathroom with a white bathtub and an open doorway leading into the adjacent room. The bathroom is decorated with black and white wallpaper, which adds a stylish touch to the space." Figure 4 on the left is generated using the Qwen-Image model. One can observe that there are two doors, which is an obvious mistake. On the right, the image is generated using the Our Full Model, which produces an image that makes the most sense in terms of spatial layout.

In Figure 5, the dynamic process of Our Full Model in generating multiple styles of building images under different denoising steps is shown. The experimental results show that the model achieves a good balance between generation efficiency and generation quality: only at the 15th step, it can generate structurally complete and stylized architectural images, and its visual effect is close to the convergence state.

Figure 4. An Example of Image Generation for the Qwen-Image model (left) and Our Full Model (right).

Figure 5. Our Full Model under Different Cue Words and with Step Image Generation.

Meanwhile, in order to further demonstrate the detailed differences of the images generated by the domain adaptive fine-tuning of Qwen-Image + FT under the same set of text prompts, the results are shown in Figure 6. It can be seen that, although the overall composition and content of the images generated by the model in step 15 and step 25 are basically the same, there are obvious differences in the detailed performance. The results generated in step 25 show better visual rationality in multiple dimensions: the natural lighting effect and pillow fabric texture in the first row of images are more realistic; the second and third rows of images are better than the results in step 15 in terms of spatial layout, size ratio of the furniture, and the overall spatial utilization of the scene, which shows the effect of more iteration steps on the improvement of image detail quality.

In order to further assess the overall advantages of the proposed method, Figures 7 and Figures 8 show the generation results of Our Full Model (left) and Qwen-Image + FT (right) side by side at the same number of sampling steps. It can be clearly observed by visual comparison that the left column of images significantly outperforms the right baseline model in terms of overall quality. In Figure 7, the first row of right-side images shows misalignment and distortion of the door frame structure, while the second row of right-side images has obvious color distortion. The results in Figure 8 further confirm this trend: the windows in the first row of right-side images are unnaturally tilted; the roof generated in the second row of right images is completely external to the building, which does not correspond to the semantic requirement of the cue word "a modern living room".

Figure 6. Qwen-Image + FT different cue words and step image generation.

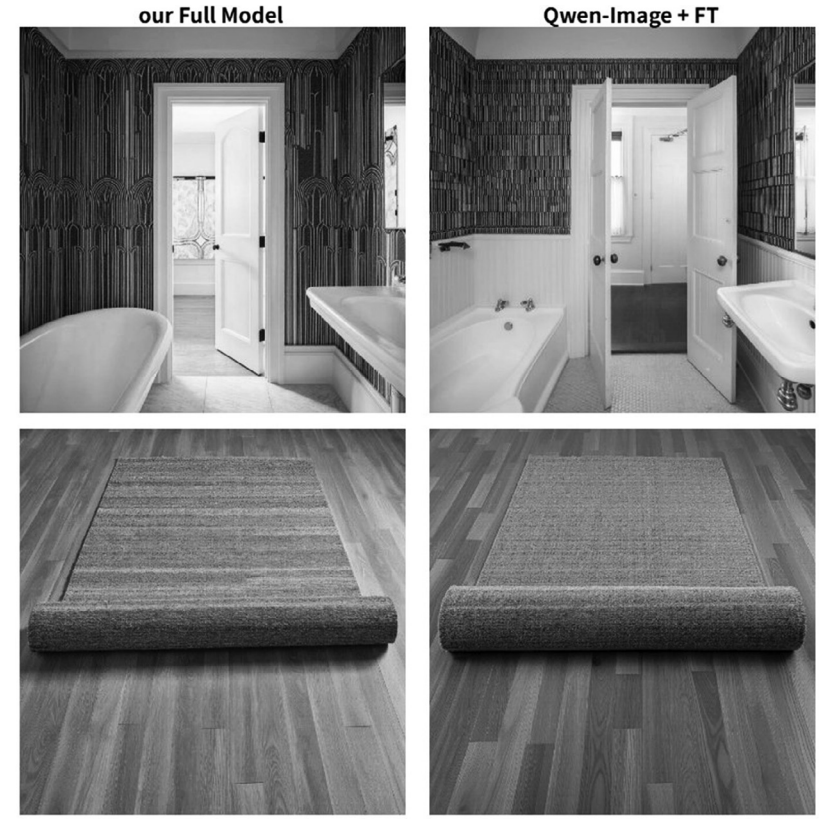

Figure 7. Our Full Model (left) and Qwen-Image + FT (right) generated building plans (1).

Figure 8. Our Full Model (left) and Qwen-Image + FT (right) generated building plans (2).

By comparing the above multi-group and multi-angle generation results, it can be concluded that Our Full Model proposed in this thesis outperforms Qwen-Image and Qwen-Image + FT model on the architectural image generation task, which verifies the effectiveness of the introduced structural consistency loss and domain adaptive strategy.

5. Discussion

The innovation of this study is that the proposed loss of structural coherence can provide new ideas for the application of Vincentian diagram models in structured domains such as architecture. In the generalized Vincentian diagram model, the attention map tends to spread over semantically similar texture regions, resulting in a high degree of overlap in the attention regions of architectural components, such as "windows" and "walls", and thus generating adherent and disproportionate components. The loss function introduced in this study allows the spatial focus loss of building components in textual cues to mimic the requirement of clarity of component positioning in architectural design by maximizing the distribution of attention and contracting the attentional quality of each building component towards a certain spatial focal point; the semantic alignment loss minimizes the cosine similarity between different components by applying an attentional vector to the different components in the representational space to ensure that the cosine similarity between "window" and "wall" is minimized. "Doors" and "windows", "roofs", and "foundations" are visually separated from each other, thus occupying a more distinguishable position in the generated image. The "door" and "window", "roof" and "foundation" are separated from each other in the visual representation, thus occupying a more distinguishable spatial region in the generated image.

The difference between our approach and the existing technology is that we do not use external methods such as ControlNet, T2I-Adapter, *etc.* to assist in image generation, nor do we optimize the text-image alignment of the model by LoRA fine-tuning alone. Rather, the loss of structural coherence was introduced under LoRA fine-tuning to allow the model to accurately localize the position and space of building-related words, which further intervened the model's spatial compositional ability more accurately through attentional correction.

However, several limitations still exist in this study. First, the structural control effect relies to a certain extent on the predefined architectural terminology dictionary, and the control of new components or descriptors that are not included may be weakened; second, although the overall structural reasonableness is improved, the control of fine-grained attributes such as furniture materials and light and shadow details is still not fine enough; third, it is currently limited to 2D image generation, and has not yet been extended to 3D spatial layout, which is still a distance away from the 3D modeling of the real architectural design process. The demand for 3D modeling in the real architectural design process is still a distance away. Based on this limitation, this method can be optimized and developed in the future.

6. Conclusion

This study proposes an improved method based on the Qwen-Image model for the precise control of style and layout in architectural text-to-image generation. The generic generative model will have problems such as chaotic layout, disproportion and inconsistent style when dealing with architectural texts containing complex spatial relations and specialized semantics. For this reason, effective control over the stylistic consistency, spatial rationality and semantic relevance of the generated images is achieved by introducing the loss of structural consistency in the model and combining it with LoRA fine-tuning, which guides the model to extract and follow the architectural structure from within the textual descriptions. This approach enhances the understanding of the structured semantics of the professional domain while maintaining the flexibility of generation and provides an idea for the practicalization of the text-generated image technology in the field of strong structural requirements such as architecture and interior design.

It is shown through experiments that this method significantly outperforms mainstream baseline models in several dimensions. In MMIS

data, Our Full Model excels in evaluating the quality metrics of the generated images, which signifies a significant improvement in the visual realism and diversity of the generated images. More importantly, it performs well in CLIP-Score, LPIPS, *etc.*, which confirms the unique advantages of the proposed method in terms of fine-grained semantic alignment and geometric structure fidelity. The ablation study further validates the effectiveness of each of the two proposed loss functions and their synergistic effects.

Future research work will be carried out in the following areas:

1. exploring the fine control of details such as furniture materials and lighting effects.
2. extending the generation of 2D architectural images to 3D interior scenes, realizing the complete design process from text to 3D space.
3. developing a user-interactive interface that allows designers to adjust text prompts in real time and view the generated effects immediately.

Declaration of Competing Interests

The authors declare no conflict of interest.

Funding

Ths research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Data used in this article are openly available at: https://github.com/AhmedMahmoudMostafa/MMIS

References

[1] M. Fisheret *et al.*, "Activity-centric Scene Synthesis for Functional 3D Scene Modeling", *ACM Trans. Graph.*, vol. 34, no. 6, p. 179, 2015. https://doi.org/10.1145/2816795.2818057

[2] R. Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models", in *Proc. of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA*, 2022, pp. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[3] K. Wang *et al.*, "Deep Convolutional Priors for Indoor Scene Synthesis", *ACM Trans. Graph.*, vol. 37, no. 4, p. 70, 2018. https://doi.org/10.1145/3197517.3201362

[4] L. Zhang *et al.*, "Adding Conditional Control to Text-to-Image Diffusion Models", in *Proc. of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France*, 2023, pp. 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355

[5] C. Mou *et al.*, "T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models", in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4296–4304. https://doi.org/10.1609/aaai.v38i5.28226

[6] H. Chefer *et al.*, "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models", *ACM Trans. Graph.*, vol. 42, no. 4, p. 148, 2023. https://doi.org/10.1145/3592116

[7] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks", in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pp. 1316–1324, 2018. https://doi.org/10.1109/CVPR.2018.00143

[8] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers", in *Proc. of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France*, 2023, pp. 4172–4182. https://doi.org/10.1109/ICCV51070.2023.00387

[9] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision", *ICML 2021*, pp. 8748–8763, 2021. https://doi.org/10.48550/arXiv.2103.00020

[10] C. Wu *et al.*, "Qwen-Image Technical Report", ArXiv. https://doi.org/10.48550/arXiv.2508.02324

[11] C. Schuhmann *et al.*, "LAION-5B: An Open Large-scale Dataset for Training Next Generation Image-text Models", *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, article no. 1833, pp. 25278–25294, 2022. https://doi.org/10.48550/arXiv.2210.08402

[12] B. Peng *et al.*, "ControlNeXt: Powerful and Efficient Control for Image and Video Generation", ArXiv. https://doi.org/10.48550/arXiv.2408.06070

[13] J. Chen *et al.*, "PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models", ArXiv. https://doi.org/10.48550/arXiv.2401.05252

[14] G. Couairon *et al.*, "Diffedit: Diffusion-based Semantic Image Editing with Mask Guidance", in *Proc. of the 11th International Conference on Learning Representation (ICLR)*, 2023. https://doi.org/10.48550/arXiv.2210.11427

[15] Y. Li *et al.*, "GLIGEN: Open-Set Grounded Text-to-Image Generation", in *Proc. of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada*, 2023, pp. 22511–22521. https://doi.org/10.1109/CVPR52729.2023.02156

[16] J. E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models", in *Proc. of the 10th International Conference on Learning Representations (ICLR)*, 2022. https://doi.org/10.48550/arXiv.2106.09685

[17] H. Kassab *et al.*, "MMIS: Multimodal Dataset for Interior Scene Visual Generation and Recognition", in *Proc. of the 2024 Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt*, 2024, pp. 172–177, 2024. https://doi.org/10.1109/IMSA61967.2024.10652794

[18] T. Salimans *et al.*, "Improved Techniques for Training GANs", *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242, 2016. https://doi.org/10.48550/arXiv.1606.03498

[19] M. Heusel *et al.*, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6629–6640, 2017.

[20] M. Bińkowski *et al.*, "Demystifying MMD GANs", in *Proc. of the 6th International Conference on Learning Representations (ICLR)*, 2018. https://doi.org/10.48550/ARXIV.1801.01401

[21] J. Hessel *et al.*, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. https://doi.org/10.48550/ARXIV.2104.08718

[22] R. Zhang *et al.*, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018 pp. 586–595. https://doi.org/10.1109/cvpr.2018.00068

Contact addresses:
YuanShuai Lan
School of Electronic Information Engineering
Geely University
Chengdu
Sichuan
China
e-mail: 448916030@qq.com

Min Liao
School of Electronic Information Engineering
Geely University
Chengdu
Sichuan
China
e-mail: liaomin@guc.edu.cn

Mo Chen
School of Electronic Information Engineering
Geely University
Chengdu
Sichuan
China
e-mail: chenmo@guc.edu.cn

Yi Ou
School of Electronic Information Engineering
Geely University
Chengdu
Sichuan
China
e-mail: ouyi@guc.edu.cn

YuanShuai Lan obtained his master's degree from Chengdu University of Information Technology, China, in 2022. He is a full-time lecturer at the School of Electronic Information Engineering, Geely University, Sichuan, China. His research mainly focuses on artificial intelligence applications, with a particular emphasis on the real-world applications of computer vision and large models.

Min Liao obtained her master's degree from Sichuan University, China, in 2018. She is a full-time faculty member at the School of Electronic and Information Engineering, Geely University of China, Sichuan. Her research focuses on artificial intelligence applications, with an emphasis on the practical integration of AI technologies in electronic and information engineering scenarios.

Mo Chen obtained her master's degree in Applied Mathematics from TU Clausthal, Germany, in 2017. Currently, she works as a full-time faculty member at Geely University of China. Her research interests focus on embedded artificial intelligence and optimization algorithms.

Yi Ou graduated from Beihang University with a master's degree in 2013. She holds the title of lecturer at the School of Electronic Information Engineering, Geely University, Sichuan, China. She specializes in data algorithms and artificial intelligence.

Received: October 2025
Revised: December 2025
Accepted: December 2025