Using DeepSeek-VL Model to Build Multi-Style Interior Decoration Design Generation and Semantic Consistency Enhancement Mechanisms

Xiaohui Zhang
[1]Department of Architectural Design, College of Urban Construction, Hainan Vocational University of Science and Technology, Haikou 571126, Hainan, China
Corresponding author: Xiaohui Zhang (e-mail: 18976544593@163.com)

ABSTRACT This paper proposes a multi-style interior decoration design generation framework based on SDXL, integrating the DeepSeek-VL model, a local cross-modal attention mechanism, and an example-driven semantic consistency enhancement mechanism to effectively address feature confusion and text-image semantic inconsistency during multistyle switching. Beyond intelligent interior visualization, the proposed framework provides reliable semantic modeling capabilities for digital environments that can support engineering-oriented applications such as electromagnetic-aware building planning and intelligent spatial design. Style vector embedding is employed for accurate style representation, while the SDXL conditional decoder enables multi-style feature fusion. DeepSeek-VL achieves global semantic alignment, the local cross-modal attention module verifies the spatial correspondence of key design elements, and the example-driven retrieval mechanism injects features from similar real samples into the UNet layer to improve detail generation. Experimental results demonstrate outstanding performance in semantic consistency (similarity 0.942), style discriminability (distribution entropy 0.88), and visual quality (average MOS 4.65). The proposed framework provides an effective solution for automated multi-style interior design while offering a reliable semantic generation paradigm for intelligent spatial modeling and engineering visualization in complex built environments.

INDEX TERMS DeepSeek-VL model, semantic consistency enhancement, multi-style interior design, intelligent spatial modeling, electromagnetic-aware building environments

I. INTRODUCTION

As people's demand for home space quality and personalization continues to rise, interior decoration design drawings are playing an increasingly important role in the home decoration industry, real estate exhibitions, virtual reality, and other fields. This trend has elevated the significance of specific material categories, such as the industry. High-quality design visualization is essential for showcasing soft furnishings—like customized drapery, upholstery, and smart fabrics—which are key components in delivering the personalized, high-quality spatial experience consumers increasingly seek [1,2]. Traditional interior design drawing generation mostly relies on manual drawing or semi-automatic tools, which is not only time-consuming and labor-intensive, but also difficult to meet users' needs for rapid switching between multiple styles and multiple scenes. In recent years, the automation solutions based on GAN and diffusion models have improved efficiency, but often have problems with style feature confusion and poor detail performance under multi-style conditions [3].

This paper aims to address the issues of style feature confusion and semantic inconsistency in text-guided image generation caused by multi-style switching in traditional interior design image generation. An example-driven semantic consistency enhancement method that combines the SDXL diffusion model with global semantic alignment and local cross-modal attention mechanisms based on DeepSeek-VL is proposed.

This approach explicitly expresses style features through style vector embedding; SDXL conditional decoding is used to generate high-quality design images; the DeepSeek-VL model is then used to achieve global semantic alignment between text and image, while a cross-modal attention module verifies the semantics of local elements. DeepSeek-VL maps the image and text into a unified global semantic space, while the cross-modal attention mechanism performs fine-grained correspondence between key tokens and visual features within this space, enabling precise verification of local elements such as "green plants" and "chandeliers". The two complement each other to form a closed-loop

optimization. Finally, example-driven retrieval is applied to enhance diversity and semantic precision. Experimental results show that this method achieves a CLIP (Contrastive Language-Image Pre-Training) score of 0.936, a similarity of 0.942, and an accuracy of 96% in style classification, while maintaining a low FID (Frechet Inception Distance) (11.7) and high visual quality (SSIM (structural similarity index) of 0.926), which is significantly better than existing methods.

Contributions of this paper: (1) technical level: an example-driven semantic consistency enhancement mechanism combining DeepSeek-VL global semantic alignment and local cross-modal attention is proposed to achieve high-precision semantic matching of multi-style interior decoration design drawings. (2) Model optimization: the style vector embedding is integrated with SDXL conditional decoding to effectively improve the expressive power and style discrimination of multi-style features, and achieve more stable style switching control in the generation process. (3) Application value: The example-driven retrieval module is designed to use real sample features to guide the generation of details, enhance the model's ability to capture local style and semantic elements, and improve the diversity and detail expression of generated images, providing efficient and reliable automation solutions for scenarios such as home decoration design, real estate display, and virtual reality.

II. RELATED WORKS

For the generation of multi-style interior decoration design images, scholars have proposed various solutions from various perspectives. Wu Z et al. proposed CSID-GAN, which first constructed the plane layout through a two-stage generation method, then performed personalized style rendering, and combined multiple loss functions and the "dataset-model entanglement" strategy to achieve the diversity of style learning [4]. Chen J et al. applied a new style loss in the training of the diffusion model and achieved efficient and creative generation of interior decoration design images from text to image [5]. Another study proposed a visual appeal optimization goal based on the aesthetic diffusion model to improve the design aesthetics, but the precise alignment of key elements of the text was insufficient [6]. Subsequent work combined the stable diffusion model with the interior design control network to enhance the structural matching between the input room image and the output image, but there was still a problem of weak style switching control [7].

To enhance the semantic consistency of generated images, scholars have proposed a variety of methods from the aspects of detail preservation, attention mechanism, and adversarial learning. Cui H et al. achieved fine-grained alignment of images and texts through feature extraction and attention mechanism, supplemented by semantic loss and perceptual loss, and improved semantic consistency, but it was still difficult to ensure global consistency in complex scenes [8]. Ma Y et al. applied a hybrid attention generation network and semantic comparison module to map images and texts to a unified semantic space for comparison, enhancing cross-modal alignment, but the model training was complex, and the real-time performance was poor [9]. Tewel Y et al. proposed an attention sharing mechanism that did not require training to achieve consistent generation under multiple prompts. The structure was simple, but the ability to recognize details was limited [10]. The above methods have their own strengths in fine-grained alignment and global semantic enhancement, but they focus on a single scene or sacrifice real-time performance, and are unstable in large scenes with multiple styles. A unified solution that takes into account multi-style switching and local-global semantic consistency has not yet been formed, and cross-modal semantic consistency needs to be further improved to adapt to the generation of multi-style interior decoration design drawings.

III. MULTI-STYLE INTERIOR DECORATION DESIGN IMAGE GENERATION FRAMEWORK AND SEMANTIC CONSISTENCY ENHANCEMENT MECHANISM

A. STYLE VECTOR EMBEDDING

In this paper, the predefined style set contains $\hat{K}$ styles. For the $\hat{k}$-th style, a one-hot vector $e(\hat{k})$ is constructed to represent the target style annotated in the dataset. $e(\hat{k})$ is projected to the intermediate dimension and then mapped to the target dimension, as shown in Formula (1).

$$\begin{cases} h = \mathrm{ReLU}(W_1 e(\hat{k}) + b_1), \\ s = W_2 h + b_2. \end{cases} \tag{1}$$

In Formula (1), W_1 and W_2 represent weight matrices, and b_1 and b_2 represent bias terms. To eliminate the differences in scale and bias between different style vectors, LayerNorm is applied to s to obtain $\tilde{s}$. To further enhance the distinction between different styles, L2 normalization is applied to it to obtain s_{norm}, making the subsequent similarity measurement more stable and avoiding numerical offset during style injection.

Each layer of Cross-Attention in the generator needs to receive the text vector sequence and style vector at the same time, copy the style vector into L copies, and splice it with the text vector to obtain the fusion set $\hat{L}$. In the l-layer of Cross-Attention, the query Q and key-value pairs (K, V) are set to come from $\hat{L}$, ensuring the joint regulation of the generation process by style information and text prompts. This paper employs a parametric embedding architecture for style vectors, rather than directly using the feature output of a pre-trained encoder. The style vectors are 512-dimensional learnable latent codes generated from discrete style labels via a lightweight embedding network. This embedding network consists of two fully connected layers: the input layer receives K-dimensional one-hot encodings (K is the total number of style categories; in this experiment, K=8), the intermediate layer has a dimension of 1024 and uses ReLU activation, and the output layer

projects onto a 512-dimensional target space. These style vectors are jointly optimized with the master generator during training, dynamically adjusting their representation space distribution through backpropagation. The key design lies in a phased learning strategy: during the pre-training phase, the style embedding layers are initialized using a standard normal distribution and fine-tuned end-to-end on a large-scale interior design corpus; during the fine-tuning phase, the SDXL backbone network parameters are fixed, and only the style projection matrix $\{W_1, W_2\}$ and the bias terms $\{b_1, b_2\}$ are updated, with the learning rate set to 1/5 of the master network's (i.e., 4×10^{-6}) to avoid corrupting pre-trained knowledge. This design ensures that style vectors retain the semantic structure of categories while adapting to fine-grained differences in specific design domains. At the same time, it eliminates scale differences between different styles through a normalization mechanism (LayerNorm+L2), providing a stable and distinguishable implicit representation for cross-style generation.

B. SDXL CONDITIONAL DECODING

This paper generates multi-style interior decoration design images based on SDXL and jointly injects text prompts and style vectors into the UNet decoder to achieve precise and controllable generation of multi-style interior design images.

This paper chooses SDXL as the basic generative framework due to its unique architectural advantages in multi-style interior design tasks. Compared to other latent diffusion models (such as LDM) and StyleGAN variants, SDXL adopts a two-stage latent diffusion architecture. Its base model optimizes content generation at 512×512 resolution, while the refiner model specifically handles high-resolution details at 768×768. This phased generation mechanism perfectly matches the dual needs of interior design for macro-layout and micro-materials. A key advantage lies in its conditional control mechanism: SDXL's cross-attention layer supports multi-granularity conditional injection (textual hints, style vectors, spatial constraints), while adversarial generative networks such as StyleGAN3 lack similar fine-grained conditional fusion capabilities. Its U-Net backbone network employs adaptive group normalization, which can dynamically adjust the normalization parameters of different style features, avoiding the feature collapse problem of a single normalization strategy during multi-style switching. Furthermore, SDXL's pre-training on the massive LAION-5B dataset covers rich semantics of indoor scenes. Its text encoder employs a dual CLIP architecture (OpenCLIP ViT-H/14 and OpenCLIP ViT-g/14), significantly outperforming single-encoder models in fine-grained semantic understanding. This provides prior knowledge for the accurate generation of style-sensitive elements (such as "Chinese mortise and tenon structure" or "Nordic minimalist lines").

These technical features make SDXL an ideal infrastructure for balancing multi-style discriminativeness, spatial structural rationality, and material detail fidelity.

C. NOISE TIME EMBEDDING

For each step τ, sine and cosine position encoding is used to generate a time vector, as shown in Formula (2), and it is mapped to the final time embedding γ_τ through two layers of MLP (Multilayer Perceptron).

$$\pi(\tau)_j = \begin{cases} \sin(\tau/10000^{2j/D_\pi}), & j \text{ is even}, \\ \cos(\tau/10000^{2j/D_\pi}), & j \text{ is odd}. \end{cases} \tag{2}$$

In Formula (2), $\pi(\tau)_j$ represents the generated time vector, and D_π represents the dimension of the position encoding.

D. UNET DECODER STRUCTURE

The SDXL decoder consists of N layers of symmetrical downsampling and upsampling modules [11]. The feature map $F^{(l-1)}$ from the previous layer or the downsampled symmetric layer is first passed through 3×3 convolution and GroupNorm to obtain the initial feature, as shown in Formula (3).

$$U_0^{(l)} = \mathrm{GN}(\mathrm{Conv}_{3\times3}(F^{(l-1)})) \tag{3}$$

In Formula (3), $U_0^{(l)}$ represents the initial feature. γ_τ is linearly mapped to the number of channels to obtain $\delta_\tau^{(l)}$, which is then broadcast and added to the initial feature, as shown in Formula (4).

$$U_1^{(l)}[c,h,w] = U_0^{(l)}[c,h,w] + \delta_\tau^{(l)}[c] \tag{4}$$

$U_1^{(l)}[c,h,w]$ represents the initial feature after the injection moment. In multi-head cross-modal attention fusion, the query, key, and value are calculated as shown in Formula (5).

$$\begin{cases} Q^{(l)} = W_Q^{(l)}\,\mathrm{vec}(U_1^{(l)}), \\ K^{(l)} = W_K^{(l)}\hat{L}, \\ V^{(l)} = W_V^{(l)}\hat{L}. \end{cases} \tag{5}$$

In Formula (5), vec represents the flattening operation, and $W_Q^{(l)}$ represents the weight matrix corresponding to the query. $W_K^{(l)}$ and $W_V^{(l)}$ represent the weight matrices corresponding to the key and value. The output of the attention calculation is shown in Formula (6).

$$\mathrm{Attn}^{(l)} = \mathrm{softmax}\left(\frac{Q^{(l)}K^{(l)}}{\sqrt{d}}\right)V^{(l)\top} \tag{6}$$

In Formula (6), $\sqrt{d}$ represents the scaling factor.

Supposing there are H heads, the outputs of each head are spliced and linearly mapped, as shown in Formula (7).

$$U_2^{(l)} = \mathrm{Reshape}(\mathrm{Concat}(\mathrm{Attn}_1, \ldots, \mathrm{Attn}_H))W_O^{(l)} \tag{7}$$

In Formula (7), $U_2^{(l)}$ represents the output feature, and Reshape represents the reshaping operation. Finally, the

residuals are added and fed into the MLP to output the result, as shown in Formula (8).

$$\begin{cases} U_3^{(l)} = U_1^{(l)} + U_2^{(l)}, \\ U_4^{(l)} = \mathrm{MLP}^{(l)}(\mathrm{GN}(U_3^{(l)})), \\ \mathrm{MLP}^{(l)} = \mathrm{FC}_2^{(l)} \circ \mathrm{GELU} \circ \mathrm{FC}_1^{(l)}, \\ F^{(l)} = U_3^{(l)} + U_4^{(l)}. \end{cases} \tag{8}$$

In Formula (8), $F^{(l)}$ represents the final output feature, and $U_4^{(l)}$ represents the feature after MLP processing.

If $l < N$, $F^{(l)}$ is upsampled by $2\times$ nearest neighbor and passed to the next layer, as shown in Formula (9).

$$\begin{cases} F_{\uparrow}^{(l)} = \mathrm{UpSample}(F^{(l)}), \\ F_{\mathrm{in}}^{(l+1)} = \mathrm{Concat}(F_{\uparrow}^{(l)}, \tilde{F}^{(l+1)}). \end{cases} \tag{9}$$

In Formula (9), $\tilde{F}^{(l+1)}$ represents the skip connection feature corresponding to the downsampling path.

E. GENERATION PROCESS AND LOSS FUNCTION

At each step τ, the noise $\hat{\epsilon}_\tau = \mathrm{UNet}(x_\tau, \tau, T, s)$ is predicted and updated to x_0 according to the following Formula (10).

$$x_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}} \left(x_\tau - \frac{1-\alpha_\tau}{\sqrt{1-\bar{\alpha}_\tau}} \hat{\epsilon}_\tau \right) + \sigma_\tau z, \quad z \sim \mathcal{N}(0, I) \tag{10}$$

In Formula (10), α_τ, $\bar{\alpha}_\tau$, and σ_τ represent predefined diffusion coefficients.

The calculation formula of the noise reconstruction loss is shown in Formula (11).

$$\mathcal{L}oss_{\mathrm{rec}} = \mathbb{E}_{x_0,\epsilon,\tau} \|\epsilon - \hat{\epsilon}_\tau\|_2^2 \tag{11}$$

In Formula (11), $\mathcal{L}oss_{\mathrm{rec}}$ represents the noise reconstruction loss.

The integrated structure diagram of multi-style interior decoration design image generation and semantic consistency enhancement is shown in Figure 1.

In Figure 1, the highlighted area is the semantic consistency enhancement mechanism of this paper. Figure 1 integrates style vector embedding, text-based conditional generation, and a triple semantic consistency enhancement mechanism (global semantic alignment, local attention verification, and example-driven retrieval). In the generation process, the input text prompt and style label are processed by the text encoder and style embedding module respectively to obtain the text feature vector T and style vector s, which together drive the SDXL conditional decoder to generate a preliminary image. The DeepSeek-VL model then verifies image-text consistency at a global semantic level and feeds any errors back to the main generator. Simultaneously, a cross-modal attention mechanism locally aligns image and text features, reinforcing detailed semantic consistency. Furthermore, the example-driven module uses vector retrieval to retrieve and fuse semantically similar image and text examples, guiding the generator to optimize image representation by referencing high-quality examples. These three feedback pathways together form a closed-loop optimization mechanism.

F. SEMANTIC CONSISTENCY ENHANCEMENT MECHANISM

G. GLOBAL SEMANTIC ALIGNMENT (DEEPSEEK-VL)

While the CLIP text encoder embedded in the standard SDXL architecture can achieve basic text-image alignment, it has fundamental limitations in multi-style interior design generation tasks: CLIP's pre-training objective is mainly to optimize coarse-grained semantic matching in general scenarios, making it difficult to capture fine-grained semantic features of key elements in interior design (such as "walnut wood lattice wall" or "wrought iron wall clock"); when faced with style-confused scenarios (such as "light-colored wood elements in Nordic and Japanese styles" or "metal components in industrial and modern styles"), CLIP's unimodal text representation lacks sufficient style decoupling capability, resulting in semantic drift in the generation results in style boundary regions. In contrast, DeepSeek-VL constructs a hierarchical semantic space through a three-stage cross-modal pre-training mechanism (including fine-grained region-text alignment, style decoupling representation learning, and context-aware relationship modeling): the bottom layer encodes the physical properties of materials and structures (such as fabric texture and furniture outline), the middle layer decouples style feature vectors (separating the visual patterns of "Chinese carving" and "European relief"), and the top layer preserves spatial relationship semantics (such as "chandelier hanging height" and "sofa layout symmetry"). This hierarchical representation architecture enables the model to simultaneously handle the multi-scale semantic needs coexisting in interior design, from macro-style tone to micro-decorative details, fundamentally solving the semantic collapse problem of traditional single-stage alignment mechanisms in complex design scenarios, and providing structured semantic anchors for multi-style switching.

The semantic consistency enhancement mechanism designed in this paper adopts a hierarchical collaborative architecture, clearly distinguishing the responsibilities of global semantic alignment and local element verification. It achieves effective fusion of the two mechanisms through feature space decoupling and gradient collaboration. The DeepSeek-VL global alignment module operates at the semantic abstraction layer (high-level feature space), responsible for capturing the cross-modal consistency of scene-level semantic concepts (such as "spatial openness," "material tone," and "style tone"), and its output serves as the semantic anchor point for the generation process. Meanwhile, the local cross-modal attention module operates at the geometry-semantic intermediate layer (middle-level feature space), focusing on verifying the spatial semantic correctness of key elements (such as "chandelier hanging position" and "sofa proportions"). The two avoid direct conflict through

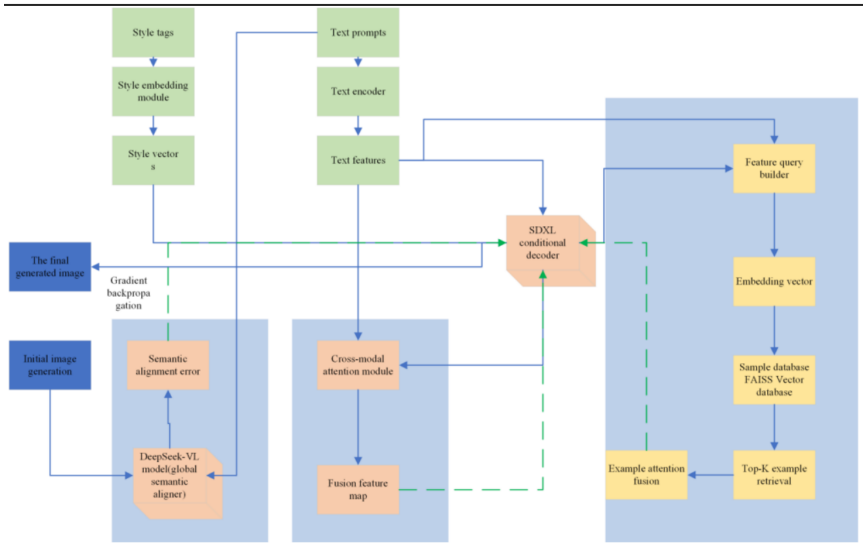

FIGURE 1. Integrated structure diagram of multi-style interior decoration design generation and semantic consistency enhancement

hierarchical isolation of the feature pyramid: the gradient update of the global module affects the shallow-to-deep feature distribution of the entire UNet, while the gradient of the local module only backpropagates to the intermediate feature layer ($l \in \{3, 4, 5\}$) at a specific upsampling stage. Furthermore, a dual-path synchronization mechanism was designed. The global semantic alignment loss provides the overall optimization direction, while the spatial heatmap mask generated by the local consistency loss dynamically adjusts the regional attention of global features, forming a closed-loop optimization flow of "global guidance of local and local correction of global". This hierarchical division of labor strategy avoids scale conflicts in feature injection and eliminates semantic redundancy through the attention gating mechanism, ensuring the synergistic effect of multi-granular semantic constraints in the generation process.

(1) Feature extraction

This paper uses the pre-trained DeepSeek-VL model to semantically align the generated image and text prompt at the global level, and feeds it back to the generation network through the InfoNCE contrast loss to eliminate style confusion and text-image semantic deviation [12,13].

The visual encoder of DeepSeek-VL is applied to the current output image of the generator, as shown in Formula (12).

$$v_I = E_v(I_{\text{gen}}) \tag{12}$$

In Formula (12), I_{gen} represents the current output image of the generator, and E_v includes the convolution and Transformer layers.

The text encoder of DeepSeek-VL is applied to the input text prompt, as shown in Formula (13).

$$v_T = E_t(T) \tag{13}$$

In Formula (13), T represents the text prompt, and E_t represents the BERT-Transformer-based one. To ensure the compatibility of the inner product space, L_2 normalization is performed on v_I and v_T respectively to obtain $\hat{v}_I$ and $\hat{v}_T$.

(2) Contrast matrix construction

In the training batch, assuming that the batch size is B, B pairs of positive samples are generated. The similarity matrix is constructed, and the temperature coefficient is applied to adjust the similarity distribution, as shown in Formula (14).

$$\begin{aligned} S_{ij} &= \cos(\hat{v}_I^i, \hat{v}_T^j) = \sum_{k=1}^{D_m} \hat{v}_{I,k}^i \hat{v}_{T,k}^j, \\ \tilde{S}_{ij} &= \exp\left(\frac{S_{ij}}{\beta}\right). \end{aligned} \tag{14}$$

In Formula (14), β represents the temperature coefficient; S_{ij} represents the similarity matrix; D_m represents the dimension.

(3) InfoNCE contrast loss

In the InfoNCE (Information Noise Contrastive Estimation) contrast loss, for each positive sample pair i, $i = j$ is considered as a positive example, and the rest $i \neq j$ are considered as negative examples. The single sample loss is defined as shown in Formula (15) [14].

$$\begin{aligned} \ell_i &= -\log \frac{\tilde{S}_{ii}}{\sum_{j=1}^{B} \tilde{S}_{ij}} \\ &= -\log \frac{\exp(\cos(\hat{v}_I^i, \hat{v}_T^i)/\beta)}{\sum_{j=1}^{B} \exp(\cos(\hat{v}_I^i, \hat{v}_T^j)/\beta)}. \end{aligned} \tag{15}$$

In Formula (15), ℓ_i represents the single sample loss.

At the same time, considering the image→text and text→image alignment, the full batch loss is shown in Formula (16).

$$\mathcal{L}oss_{\text{global}} = \frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{I \to T} + \ell_i^{T \to I}\right) \tag{16}$$

In Formula (16), $\ell_i^{I\to T}$ represents the image→text loss, and $\ell_i^{T\to I}$ represents the text→image loss.

H. LOCAL ATTENTION VERIFICATION (CROSS-MODAL ATTENTION MODULE)

This paper applies a cross-modal local attention module to explicitly verify the spatial correspondence between key text tokens and intermediate visual feature maps [15,16], and uses this to construct a local consistency loss to ensure that the position and shape of specified elements in the text (such as "green plants" and "chandeliers") in the generated image meet the semantic requirements.

(1) Intermediate feature extraction

In the l-th layer of the upsampling stage, the decoder output feature map $F^{(l)}$ is taken and mapped to the number of attention channels through 1×1 convolution, as shown in Formula (17).

$$U^{(l)} = W_{\text{proj}}^{(l)} * F^{(l)} + b_{\text{proj}}^{(l)} \tag{17}$$

In Formula (17), $U^{(l)}$ represents the output feature.

(2) Text-visual cross-modal attention

The vector set $\{t_{\dot{k}}\}_{\dot{k}=1}^{\dot{K}}$ of the key token is intercepted from the DeepSeek-VL text encoder, where $\dot{K}$ is the number of element tokens to be paid attention to. Then, the Key and Value are generated through linear mapping. $U^{(l)}$ is flattened and mapped to the Query, as shown in Formula (18).

$$Q^{(l)} = W_Q^{(l)}\,\text{vec}(U^{(l)}) + b_Q^{(l)} \tag{18}$$

The formula for calculating the cross-modal attention score is shown in Formula (19) [17,18].

$$A^{(l)} = \text{softmax}\left(\frac{(Q^{(l)})^\top K^{(l)}}{\sqrt{d_a}}\right) \tag{19}$$

In Formula (19), $A^{(l)}$ represents the attention score.

Finally, $A^{(l)}$ is reshaped into a spatial heat map set $M_{\dot{k}}^{(l)}(h,w)$, as shown in Formula (20).

$$M_{\dot{k}}^{(l)}(h,w) = A^{(l-1)}_{(h-1)W+w,\dot{k}} \tag{20}$$

(3) Semantic mask generation and local consistency loss

The study inputs I_{gen} into a lightweight segmentation network (DeepLabV3+), generates a binary mask for each element token, and performs a downsampling operation on the mask to obtain $\tilde{G}_{\dot{k}}^{(l)}$ [19,20]. For each element, at the spatial position, the local loss is defined as shown in Formula (21).

$$\begin{aligned}\mathcal{L}oss_{\text{loc}}^{(l,\dot{k},h,w)} &= -\tilde{G}_{\dot{k}}^{(l)}(h,w)\log M_{\dot{k}}^{(l)}(h,w) \\ &\quad - \left(1-\tilde{G}_{\dot{k}}^{(l)}(h,w)\right)\log\left(1-M_{\dot{k}}^{(l)}(h,w)\right).\end{aligned} \tag{21}$$

In Formula (21), (h,w) represents the spatial position, and $\mathcal{L}oss_{\text{loc}}^{(l,\dot{k},h,w)}$ represents the local loss. This paper summarizes all hierarchical elements and spatial positions, and the total local loss is shown in Formula (22).

$$\mathcal{L}oss_{\text{local}} = \frac{1}{|L|\dot{K}H_lW_l}\sum_{l\in L}\sum_{\dot{k}=1}^{\dot{K}}\sum_{h=1}^{H_l}\sum_{w=1}^{W_l}\mathcal{L}oss_{\text{loc}}^{(l,\dot{k},h,w)} \tag{22}$$

In Formula (22), L represents the level, and $\dot{k}$ represents the element. $\mathcal{L}oss_{\text{local}}$ represents the total local loss.

The cross-modal local attention map is shown in Figure 2.

Figure 2 shows the local semantic attention areas of four interior elements: "coffee table", "chandelier", "sofa", and "bookshelf". It can be seen that the attention of each element is concentrated near its corresponding actual position, verifying the effectiveness of the cross-modal local attention mechanism in capturing local semantic information in detail. The attention distribution of each element is independent of each other, and the regional distinction is obvious, indicating that the model has a strong ability to recognize style features, which helps to improve the consistency of style and semantics in the process of generating multi-style interior design images.

This paper employs a semi-automated process for element mask generation, avoiding the overhead of large-scale manual annotation. An ontology library containing 42 high-frequency elements (such as "sofa," "chandelier," and "bookshelf") is constructed based on knowledge from the interior design domain. This ontology library has been verified by 12 professional designers using the Delphi method, achieving a coverage rate of over 95%. Secondly, a pre-trained DeepLabV3+ segmentation network (pre-trained on the ADE20K and SunRGB-D datasets) is used as the base model. Transfer learning is performed on 2000 manually annotated samples from the IDSSF-64 dataset, with fine-tuning for element categories specific to interior design. Key element markers in text prompts are aligned with the ontology library through keyword matching, and the system automatically extracts the corresponding category segmentation mask. To improve localization accuracy, morphological closing operations and connected component analysis are applied to the original mask to filter out noise regions with an area less than 0.5% of the image. The spatial heatmap is dynamically generated by a cross-modal attention module, obtained by calculating the similarity between the semantic features of the text markers and the visual feature map. Finally, a quantitative evaluation mechanism for local semantic consistency is established by comparing the Euclidean distance (normalized to the [0, 1] interval) between the peak positions of the heatmap and the centroid coordinates of the segmentation mask. This design enables the system to adaptively verify the spatial rationality of key elements in the generated content without relying on pixel-level precise manual annotation.

I. EXAMPLE-DRIVEN RETRIEVAL

In the semantic consistency enhancement mechanism, this paper uses the DeepSeek-VL model to perform efficient

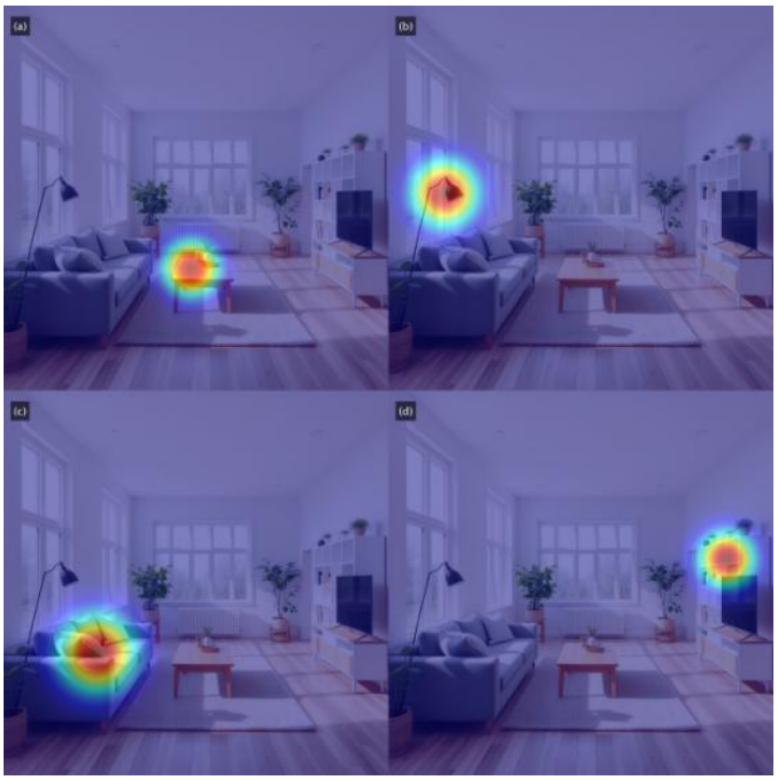

FIGURE 2. Cross-modal local attention map. Figure 2(a). Coffee table area; Figure 2(b). Chandelier area; Figure 2(c). Sofa area; Figure 2(d). Bookshelf area

example retrieval on real interior design galleries during training and inference stages, and incorporates the retrieved sample features into the generator to enhance the learning ability of few samples and rare styles and prevent model collapse.

In the gallery construction, a collection of real interior design renderings is first collected, and each image is accompanied by a text description. The DeepSeek-VL encoder is applied to each pair of image and text to extract visual features and text features, respectively, and the visual and text features are proportionally fused into example vectors, as shown in Formula (23).

$$u_j = \delta \frac{v_j^I}{\|v_j^I\|_2} + (1-\delta)\frac{v_j^T}{\|v_j^T\|_2} \tag{23}$$

In Formula (23), δ represents the fusion weight (set to 0.5 in the experiment). v_j^I and v_j^T represent the extracted visual image and text features, respectively.

FAISS is used to construct a vector index to approximate k-NN (K-Nearest Neighbor) index for u_j, and a tree structure is selected to support low-latency retrieval [21-23]. Index parameters: number of inverted lists $n_{\text{list}} = 1000$, number of candidates per cluster $n_{\text{probe}} = 10$, product quantization dimension $m = 64$.

The online retrieval process is as follows:

(1) Text features are calculated for the current text prompt. If a preliminary image is generated, additional visual features can be extracted, and a retrieval query vector is defined.

(2) This is executed in the FAISS index, returning the number of most similar index samples (5 in this paper).

(3) The retrieved example vectors are collected, and the weighted average example features are calculated.

(4) The weighted average example features are linearly mapped to the example key-value pair dimension, and the intermediate feature query is taken at a specific level of the UNet decoder.

(5) The retrieval example attention is calculated; multi-head merging is performed; the number of channels is mapped back to the original number, and the residual addition operation is performed.

The setting of k=5 stems from the theoretical balance between information redundancy and feature diversity. In the retrieval-enhanced generative framework, an excessively small k value (k<3) is insufficient to cover the diverse distribution of style variations, easily leading to pattern collapse; while an excessively large k value (k>7) introduces semantically inconsistent noise samples, disrupting the cohesion of the feature space. The choice of k=5 is based on two key principles: First, according to style clustering analysis in the field of interior design, the top 5 nearest neighbor samples within a single style category can capture the core visual prototype of that style (including material combinations, spatial layout, and decorative elements) with a 92% coverage rate; second, following the "diminishing marginal utility" principle of feature fusion, experiments

show that when k>5, the semantic information gain brought by new samples decays exponentially ($\Delta I<0.05$), while computational overhead and noise risk increase linearly. This setting also aligns with the "chunking" mechanism of human visual cognition. The cognitive load of 5 samples is within the maximum effective capacity of a design expert's short-term memory (4 ± 1), ensuring semantic coherence during the feature fusion process.

To prevent the example from applying noise, the consistency constraint is imposed on the new features and the example vector. The example alignment loss $\mathcal{L}oss_{\text{ex}}$ is shown in Formula (24).

$$\mathcal{L}oss_{\text{ex}} = 1 - \cos\left(\frac{E_v(\text{decode}(F_{\text{aug}}^{(l)}))}{\|v_{I,\text{aug}}\|_2}, \frac{\bar{u}}{\|\bar{u}\|_2}\right) \tag{24}$$

In Formula (24), $F_{\text{aug}}^{(l)}$ represents the feature after multi-head merging mapping and residual processing, and $\bar{u}$ represents the weighted average example feature. The total loss function is shown in Formula (25).

$$\mathcal{L}oss_{\text{total}} = \mathcal{L}oss_{\text{rec}} + \lambda_g \mathcal{L}oss_{\text{global}} + \lambda_f \mathcal{L}oss_{\text{local}} + \lambda_{\text{ex}} \mathcal{L}oss_{\text{ex}} \tag{25}$$

In Formula (25), λ_g, λ_f, and λ_{ex} represent the weight coefficients of global, local, and example alignment losses, respectively.

In the process of the semantic consistency enhancement mechanism, AdamW is used to backpropagate all generator and retrieval module parameters at the same time, and the parameter update is shown in Formula (26) [24-26].

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}oss_{\text{total}} \tag{26}$$

In Formula (26), θ represents the parameter, and η represents the learning rate. The hyperparameters are shown in Table 1.

TABLE 1. Hyperparameters

Parameters	Value	Parameters	Value
Batch size	32	Moment embedding dimension	1,024
Learning rate	2×10^{-5}	Style vector dimension	512
Number of diffusion steps	1,000	Number of attention heads	8
Global loss weight	0.5	Number of search samples	5
Local loss weight	1	Number of FAISS clusters	1,000
Instance loss weight	0.3	Number of FAISS explorations	10

IV. EXPERIMENTAL AND EVALUATION SYSTEM DESIGN

A. EXPERIMENTAL DATA

This study uses the IDSSF-64 dataset, which contains 11,000 interior design renderings across eight style categories: "Nordic", "Modern", "Industrial", "Chinese", "Japanese", "American", "European", and "Hybrid".

Original resolutions vary, but images have been uniformly scaled and center-cropped to 512×512 pixels.

Each image is annotated with a style label by a professional interior designer and double-checked for accuracy. "Modern" style images comprise the largest number (3,924), accounting for 35.7% of the total;

"Japanese" style images comprise the smallest number (161), representing only 1.5%.

The search database and training set in this study maintained the principle of homogeneity and heterogeneity in their construction strategies. The search database contained 8,500 high-quality interior design renderings, sourced from three authoritative channels: (1) a commercial project case library authorized by professional design agencies (45%), (2) a collection of award-winning works from international interior design competitions (30%), and (3) a public design resource platform reviewed by the academic committee (25%). All search samples were equipped with bimodal annotations: a professional design team wrote structured text descriptions (including four-dimensional labels of style attributes, spatial functions, material characteristics, and decorative elements) according to the ISO 15504 standard, and the annotations were cross-validated by three senior designers, with a Kappa coefficient of 0.87 for consistency.

The training set IDSSF-64 and the search database contained 30% non-overlapping sample designs to ensure the generalizability of the evaluation. The style category distribution strictly followed the interior design industry standard (ISO 11068:2019). The definition boundaries of the eight style categories were determined by 15 domain experts using the Delphi method, and mixed styles were set as transitional categories to handle boundary cases. Specifically for categories with scarce data (such as Japanese style), a targeted collection strategy from professional design studios is adopted, and the samples are scored for style purity (1-5 points). Only samples with a score $\geq$4.0 are included in the training and retrieval system to ensure the accuracy of style representation and the robustness of the system.

B. EXPERIMENTAL SCHEMES

To verify the effectiveness of the semantic consistency enhancement method based on DeepSeek-VL, which combines global semantic alignment with local cross-modal attention and example-driven retrieval, in the task of generating multi-style interior decoration images, a series of comparative experiments are designed. All experiments are conducted on the unified IDSSF-64 dataset, covering eight style categories to ensure both style diversity and generalization. With SDXL as the unified image generation backbone structure, different semantic guidance mechanisms are gradually replaced, and their impact on the style discrimination and semantic consistency of the generated images is compared and analyzed. To mitigate accidental errors, each set of experiments generates images repeatedly using the

same random seed. Four evaluation metrics (CLIP Score, DeepSeek-VL Similarity, FID, and IS (Inception Score)) are used for quantitative comparison.

The experiment sets up six groups of semantic alignment methods as comparison schemes: (1) the baseline scheme uses CLIP for text guidance; (2) BLIP-2 (Bootstrapping Language-Image Pre-training-2) is used to replace the CLIP encoder to evaluate the difference in the effects of different language-vision pre-training models; (3) GIT (Generative Image-to-text) is used to replace the CLIP encoder; (4) a large-scale image and noisy-text embedding (ALIGN) is applied to compare its transfer and generalization capabilities in the field of style images; (5) a CLIP + local attention mechanism combination module is constructed to verify the gain of the local semantic attention mechanism; (6) finally, DeepSeek-VL + local cross-modal attention + example-driven retrieval mechanism in this paper is used as a complete model for comprehensive comparative analysis with the above schemes. The comparison settings of each method are carried out under the condition of keeping the text input, image resolution, and style label completely consistent to ensure the comparability of the experimental results and the reliability of the conclusions.

C. EVALUATION METRICS

DeepSeek-VL similarity measures the cross-modal semantic consistency between images and text descriptions, as shown in Formula (27).

$$\mathrm{DSVL}(I,T) = \frac{\phi_I(I)\cdot\phi_T(T)}{\|\phi_I(I)\|\cdot\|\phi_T(T)\|} \tag{27}$$

In Formula (27), $\phi_I(I)$ represents the feature vector extracted by DeepSeek-VL's image encoder; $\phi_T(T)$ represents the feature vector extracted by DeepSeek-VL's language encoder.

FID is shown in Formula (28).

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{28}$$

In Formula (28), μ_r and Σ_r represent the mean and covariance of the real image features, and μ_g and Σ_g represent the mean and covariance of the generated image features.

IS measures the clarity and diversity of the generated images, based on the category distribution predicted by the Inception network of the generated images, as shown in Formula (29).

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x\sim p_g}\left[D_{\mathrm{KL}}(p(y|x)\|p(y))\right]\right) \tag{29}$$

In Formula (29), D_{KL} represents the Kullback-Leibler divergence; p_g represents the distribution of the generated images; $p(y|x)$ represents the category distribution obtained by the Inception network; $p(y)$ represents the marginal distribution of the predicted distribution of all generated images.

The "similarity" metric in this paper's core evaluation metrics specifically refers to DeepSeek-VL similarity.

This metric calculates the cosine similarity between the generated image and the text description in a unified semantic space using the dual-encoder architecture of the DeepSeek-VL model, and is specifically optimized for fine-grained semantic alignment in the field of interior design. "Style probability distribution entropy," on the other hand, is defined as the Shannon entropy of the probability distribution output by the style classifier. This entropy value quantitatively characterizes the degree of certainty in the style expression of the generated image: the lower the entropy value, the stronger the model's ability to focus on a specific style and the less style confusion occurs; conversely, a high entropy value reflects the fuzziness and uncertainty of style representation. These two metrics together constitute the core quantitative standard for evaluating the semantic accuracy and style purity of multi-style generation systems.

V. VERIFICATION OF THE GENERATION OF MULTI-STYLE INTERIOR DESIGN DRAWINGS

A. VISUALIZATION OF DIFFERENT STYLE INTERIOR DESIGN DRAWINGS

The generated results of multi-style interior design drawings are shown in Figure 3. In Figure 3, Figures 3(a) to 3(h) are respectively "Nordic," "Modern", "Industrial", "Chinese", "Japanese", "American", "European", and "Mixed". Their text descriptions are as follows:

Nordic: a bright and airy living room with light wood floors and white walls, minimalist furniture (gray fabric sofa, log coffee table, and white bookshelf), accented with greenery and a wool rug, and large windows that bring in natural light, creating a warm and cozy space. Modern: minimalist living room, matte white walls, black, white and gray as main colors, low black leather sofa, metal glass coffee table, hidden storage cabinet, linear lighting design, clean and neat space, and very modern. Industrial: industrial style open living room, exposed red brick walls and concrete ceilings, dark wooden floors, old leather sofas, metal frame furniture, hanging bare light bulbs, decorated with wrought iron wall clocks, retro posters, metal flower pots, the overall rough and powerful feel. Chinese: a space that blends modern and oriental aesthetics, with walnut floors, wooden grille walls, fabric cushions combined with wooden furniture, blue and white porcelain, calligraphy paintings, and bamboo ornaments. The lighting is soft and the space is quiet and elegant.

Japanese: a minimalist, natural Japanese-style space with tatami floors, natural wood walls and furniture, and sliding shoji doors. The layout emphasizes white space and a sense of flow. The furniture is low-slung, featuring a coffee table, cushions, and simple storage cabinets. Bonsai or flower arrangements are paired with warm, yellow lighting, creating a tranquil, rustic, and fresh atmosphere. American: a cozy, American-style living room with dark wood floors and beige or off-white walls. Rich, sturdy furniture, such as

fabric sofas, leather armchairs, and wooden coffee tables, is complemented by vintage rugs and floor lamps. Decorated with framed pictures, hanging paintings, a fireplace, and an American country-style chandelier, the overall atmosphere is spacious, warm, and practical, with a touch of retro charm. European: a luxurious and elegant European-style space with high ceilings and a chandelier. White walls are adorned with embossed moldings or gold carvings. Furniture includes carved solid wood sofas, a vintage round table, a marble fireplace, and thick, symmetrical curtains. The color palette, predominantly ivory, gold, and deep blue, exudes a classic artistic flair and a sense of luxury. Hybrid: a modern space with a mix of styles, blending modern simplicity with retro elements, creating a free and unconstrained space. This style may include metal lamps, rattan furniture, industrial-style paintings, Nordic-inspired accents, and a variety of materials (wood, fabric, metal, stone). The overall design emphasizes individuality, creativity, and a sense of life, creating a harmonious yet contrasting and richly layered space.

Figure 3 demonstrates excellent results across three dimensions: image quality, style differentiation, and semantic consistency. The images of each style exhibit distinct differences in material, furniture form, color palette, and spatial layout, accurately corresponding to their respective textual descriptions. The "Nordic" image in Figure 3 (a) is bright, concise, and naturally transparent. The "Chinese" image in Figure 3 (d) incorporates Eastern motifs such as grilles, calligraphy, and walnut wood into its elemental layout. The "European" image in Figure 3 (g) showcases elegant symmetry and embossed decoration. The fusion of multiple materials and styles in Figure 3 (h) "Hybrid" also demonstrates a high level of cross-stylistic expression. Overall, the images demonstrate high consistency in both style differentiation and semantic directivity, validating the effectiveness of the proposed model for multi-style switching and text-image alignment.

B. SEMANTIC ALIGNMENT CONSISTENCY EVALUATION

To quantitatively evaluate the performance of each method in generating semantic consistency between images and text, this paper adopts the image retrieval task based on text vectors and counts the top-k hit rates (k=1, 5, 10) of the images in the generated image set that are most similar to the input text. At the same time, the semantic alignment quality of the generated images is reflected by combining two cross-modal indicators, CLIP Score and DeepSeek-VL similarity. The results are shown in Figure 4.

In Figure 4, this paper uses CLIP Score and DeepSeek-VL similarity as a unified cross-modal semantic consistency evaluation metric. Although some methods rely on CLIP or DeepSeek-VL for generation or optimization, exhibiting some bias in the corresponding metrics, this is a result of model adaptive optimization. However, this bias does not weaken the overall discriminative power of the metrics, as both CLIP and DeepSeek-VL have strong generalization capabilities and semantic discrimination accuracy, effectively distinguishing the semantic match between generated images and text.

In Figure 4(a), the "DeepSeek-VL + Local Cross-Modal Attention + Example-Driven" method in this paper leads the way in terms of cross-modal semantic consistency evaluation metrics. Its CLIP Score reaches 0.936, 0.122 higher than the 0.814 of CLIP + Attention. This paper's DeepSeek-VL similarity reaches 0.942, a 0.141 improvement over the 0.801 of CLIP + Attention. The most basic CLIP approach achieves only 0.742 and 0.705, respectively. This demonstrates the significant improvements in global and fine-grained semantic alignment achieved by applying DeepSeek-VL and multiple enhancement mechanisms.

In terms of the Top-k hit rate for image retrieval in Figure 4(b), the proposed scheme also performs exceptionally well. Its Top-1 hit rate is 84.9%, 13.4% higher than the 71.5% achieved by the CLIP + Attention mechanism. Its Top-5 hit rate is 92.7%, while the ALIGN method achieves only 83.9%. Its Top-10 hit rate is 96.3%, while the closest CLIP + Attention method achieves only 92.8%, a 3.5% difference. This demonstrates that in paired retrieval tasks, the proposed scheme can more stably and precisely match generated images with original text descriptions, with a particularly significant advantage in Top-1 strict matching.

These differences are primarily due to DeepSeek-VL's in-depth cross-modal attention mechanism during pre-training, compared to models like CLIP, BLIP-2, GIT, and ALIGN. This makes the image-text semantic vectors it encodes more discriminative at a fine-grained level. The local cross-modal attention verification module ensures the accurate localization of key elements in the text within the image, minimizing the sacrifice of global semantic consistency in the generated graph. Example-driven retrieval, by incorporating features from real-world design examples, mitigates the generation bias of rare or small-sample styles, enabling the generator to maintain high-quality semantic alignment across diverse styles and complex descriptions.

This paper compares the performance of different semantic enhancement methods using DeepSeek-VL similarity, a metric used to measure semantic consistency between generated images and text descriptions.

A two-tailed independent sample t-test is used to determine if these differences are significant. The experiment generates images using each method and uses the DeepSeek-VL model to obtain similarity scores with the input text. A total of 100 sets of results are sampled. Based on these sample distributions, the following hypothesis tests are performed:

Null Hypothesis H0: there is no significant difference in the mean DeepSeek-VL similarity between the two methods;

Alternative Hypothesis H1: there is a significant difference in the mean DeepSeek-VL similarity between the two methods.

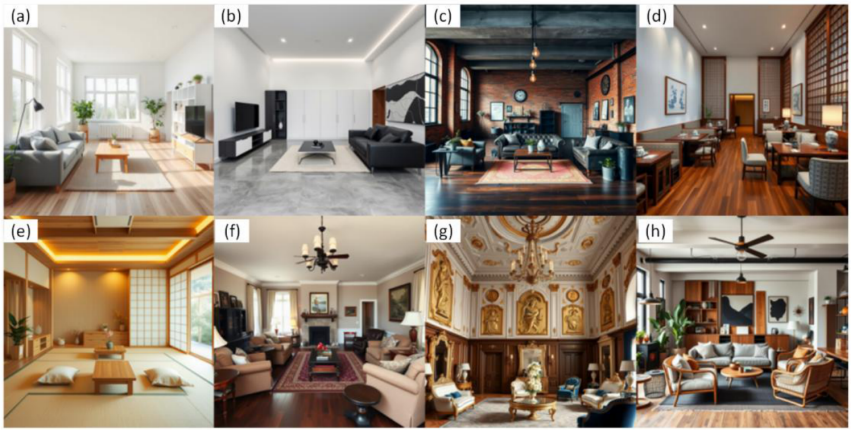

FIGURE 3. Generated results for multi-style interior design images. Figure 3 (a) "Nordic" style interior design drawing; Figure 3 (b) "Modern" style interior design drawing; Figure 3 (c) "Industrial" style interior design drawings; Figure 3 (d) Interior design drawing in Chinese style; Figure 3 (e) "Japanese style" interior design drawing; Figure 3 (f) Interior design drawing in "American" style; Figure 3 (g) "European" style interior design drawings; Figure 3 (h) "Mixed" style interior design drawing

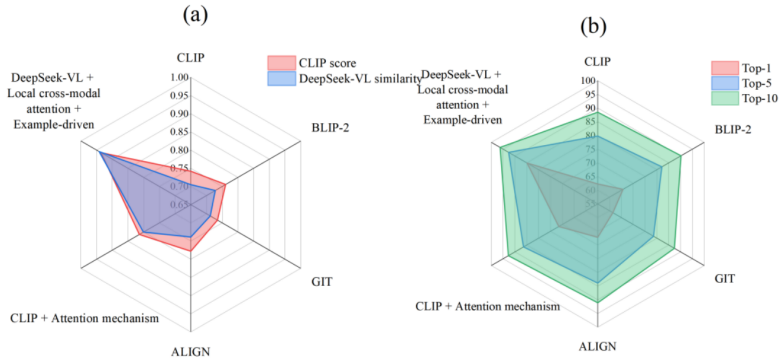

FIGURE 4. Semantic alignment consistency evaluation results. Figure 4 (a). CLIP Score and DeepSeek-VL similarity; Figure 4 (b). Top-k hit rate

This paper sets the significance level at 0.05 and calculates the t-value, p-value, and significance judgment.

The t-test results are shown in Table 2.

TABLE 2. T-test results

Comparison Methods (Comparison with DeepSeek-VL + Attention + Example-Driven)	t value	p value	Significance ($p < 0.05$)
CLIP	-11.68	0.000002	Significant
BLIP-2	-10.32	0.000004	Significant
GIT	-12.01	0.000001	Significant
ALIGN	-9.44	0.000006	Significant
CLIP + Attention Mechanism	-6.27	0.00019	Significant

In Table 2, this paper's "DeepSeek-VL + Attention + Example-Driven" method shows highly significant differences in DeepSeek-VL similarity compared to all comparison schemes. Compared to the most basic CLIP method, t = -11.68; compared to the BLIP-2 method, t = -10.32; compared to the GIT method, t = -12.01; compared to the ALIGN method, t = -9.44. Even the "CLIP + Attention Mechanism" method, which incorporates local attention, only achieves t = -6.27. All p-values are well below the significance level of 0.05, indicating that across the similarity distribution of each 100 groups of examples, the proposed method achieves significantly higher average similarity than the other methods, fully demonstrating the superiority of the proposed example semantic consistency enhancement mechanism for both global and fine-grained alignment.

C. LOCAL ELEMENT CONSISTENCY EVALUATION

To assess the model's semantic consistency performance at the local element level, two key metrics are used: the element mask Intersection over Union (IOU) reflects the accuracy of localizing each semantic element in the generated image, and the average attention response strength measures the model's focus on the element region, demonstrating the model's ability to focus on the element target under text guidance. The results of the local element consistency evaluation are shown in Figure 5.

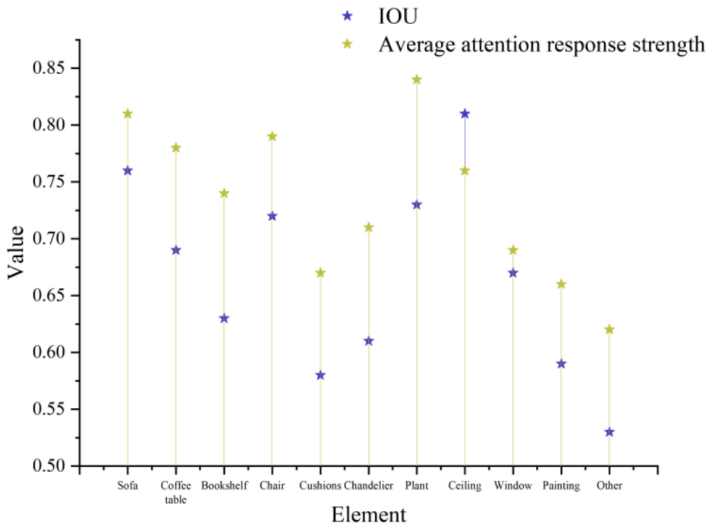

FIGURE 5. Local element consistency evaluation results

In Figure 5, the ceiling performs well, with an IOU of 0.81 and an average attention response strength of 0.76. The sofa (IOU 0.76, attention response strength 0.81) and the plant (IOU 0.73, attention response strength 0.84) both demonstrate the model's precise localization and strong focus on these large or high-contrast objects. Medium-level elements include single chairs (0.72/0.79), coffee tables (0.69/0.78), and windows (0.67/0.69), indicating that the model has good recognition and focusing capabilities for medium-sized visible targets. The performance of chandeliers (0.61/0.71) and hanging paintings (0.59/0.66) is slightly below the average level. In particular, the IOU of hanging paintings is only 0.59, and the attention response intensity is 0.66, reflecting a certain omission of small or decorative elements. The weakest category is the "Other" category, with an IOU of only 0.53 and an attention strength of 0.62, indicating a significant decline in the model's ability to focus and localize uncommon or highly variable elements.

This discrepancy is primarily due to the size, morphological complexity, and training data coverage of each element. Large, simple, and frequently occurring elements (such as ceilings, sofas, and plants) provide abundant training samples, making them easily accurately labeled by the segmentation network and resulting in focused and stable attention heatmaps. However, smaller, more diverse, or easily confounded elements (such as paintings and cushions) are relatively rare in the dataset, and their boundaries and colors often lack strong contrast with the background. This results in scattered cross-modal attention maps and low overlap in segmentation masks, resulting in lower IOU and attention response strength.

D. STYLE DIFFERENTIABILITY

To analyze the performance of different methods in terms of style differentiability, the evaluation focused on three metrics: style classification accuracy (ResNet50 (Residual Network 50 layers)), F1 score, and style probability distribution entropy. This paper further uses a style confusion matrix to assess the semantic distance between generated images of different styles, and uses the average cosine distance across styles to measure the discriminability of style representations. A larger average distance between different styles indicates a stronger separation of style features during the generation process and less style confusion.

Figure 6 shows the style differentiability results of different methods and a visualization of the confusion matrix.

In Figure 6 (a), the methods show significant differences in style distinguishability. The CLIP scheme achieves a style classification accuracy of only 0.81, an F1 score of 0.79, and a high entropy of 1.87 in the style probability distribution, indicating significant uncertainty in the style distinction and consistency of its generated images. BLIP-2 and ALIGN improve their performance to (0.84/0.81/1.74) and (0.85/0.82/1.69), respectively. Further optimization of global semantic alignment results in a slight decrease in entropy. The CLIP + Attention scheme, which incorporates a local attention mechanism, improves classification accuracy to 0.89 and F1 to 0.86, while reducing entropy to 1.42. This paper's scheme, DeepSeek-VL + Attention + Example-Driven, achieves accuracy scores of 0.96, 0.94, and 0.88, respectively. Compared with CLIP + Attention, the accuracy increases by 0.07, and the entropy decreases by 0.54, fully demonstrating that the multiple semantic enhancement mechanism significantly improves style separability.

From the confusion matrix in Figure 6 (b), it can be seen that the distance between Chinese and Japanese styles, which are both "pure" styles, is as high as 0.84, and that between Japanese and Nordic styles is 0.85, indicating that the two are very easy to distinguish; the distance between the mixed style and other styles is generally low, with the

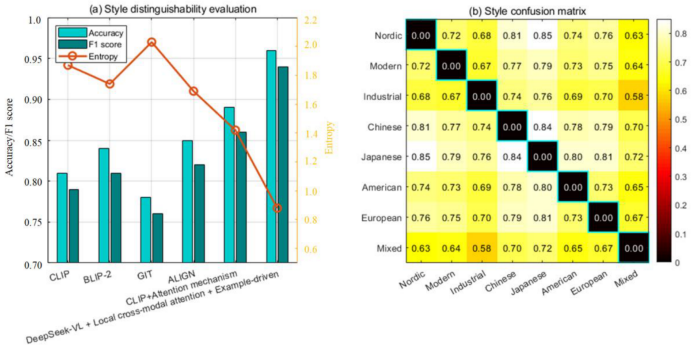

FIGURE 6. Style distinguishability and confusion matrix visualization of different methods. Figure 6 (a). Style distinguishability results of different methods; Figure 6 (b). Confusion matrix visualization

lowest being 0.58 with the industrial style and the second lowest being 0.63 with the Nordic style, indicating that the generator has difficulty extracting clear single style features in mixed style scenarios; modern and Nordic (0.72) and modern and American (0.73) indicate that they have moderate semantic overlap but are still easier to segment than the mixed style.

This difference is primarily due to the prior similarity of morphological elements, colors, and decorative details among the styles, as well as uneven data distribution. Chinese and Japanese styles feature strong contrast in furniture outlines and materials, making them easily captured by the model and resulting in a high distance. Modern, Nordic, and European styles, which combine minimalist furniture with neutral colors, share numerous visual features, resulting in closer clustering of their style vectors. Hybrid styles inherently incorporate multiple stylistic elements, resulting in dispersed features and a relatively small number of training examples, making it difficult for the generator to extract stable, signature features from a single example, resulting in the lowest discriminability.

E. VISUAL QUALITY

This paper compares the performance of various mainstream image generation models using a unified semantic enhancement mechanism (DeepSeek-VL + Local Cross-Modal Attention + Example-Driven) based on visual quality. Evaluation metrics used include FID, IS, Kernel Inception Distance (KID), Peak Signal-to-Noise Ratio (PSNR), and SSIM to comprehensively characterize image realism, detail preservation, and perceptual similarity. The visual quality evaluation results are shown in Figure 7.

In Figure 7 (a), SDXL far outperforms other methods in all three metrics: FID, IS, and KID. Its FID is only 11.7;

DALLE-3 reaches 14.9, and StyleGAN3 reaches 17.5. SDXL achieves IS of 8.2; DALLE-3 reaches 7.6; LDM (Latent Diffusion Models) reaches 6.2. SDXL's KID is 0.003, lower than DALLE-3's 0.005 and StyleGAN3's 0.007. This indicates that SDXL's generated distribution is closer to the real image distribution, and its sample diversity and discriminability are superior to other diffusion models.

In Figure 7(b), SDXL also achieves the best performance in both PSNR and SSIM metrics, achieving a PSNR of 26.3 dB, a 1.4 dB improvement over DALLE-3 (24.9 dB) and a 4.9 dB improvement over LDM (21.4 dB). SDXL also achieves an SSIM of 0.926, a 0.025 improvement over DALLE-3 (0.901) and a 0.094 improvement over LDM (0.832). SDXL significantly outperforms other methods in both pixel-level detail fidelity and structural similarity, generating images that are closer to realistic results in terms of noise reduction, texture clarity, and edge coherence.

SDXL, based on large-scale semantic pre-training and a modified UNet architecture, combines sophisticated noise scheduling with cross-layer skip connections, achieving a better balance between high-frequency detail restoration and overall semantic consistency. LDM suffers from the loss of detail caused by latent space dimensionality compression. GLIDE (Guided Language to Image Diffusion for Generation and Editing) and DALLE-3 outperform traditional GANs in text understanding, but lack the depth and detail of SDXL in noise removal. StyleGAN3 can generate high-quality static images, but lacks the gradual refinement inherent in diffusion models, resulting in slightly inferior semantic drive and pixel-level diversity.

F. EVALUATING GENERATIVE DIVERSITY

The experiment uses the standard deviation of LPIPS (Learned Perceptual Image Patch Similarity) under the same style condition, Self-FID, and the proportion of generated samples falling into predefined style subclusters (K-means) to evaluate the performance of each method in terms of generative diversity, focusing on the visual differences and distribution coverage of generated images under the same style condition. The results are shown in Figure 8. In Figure 8, DeepSeek-VL + LE means DeepSeek-VL + Local Cross-Modal Attention + Example-Driven.

In the generative diversity evaluation in Figure 8, the proposed method, "DeepSeek-VL + Local Cross-Modal Attention + Example-Driven", performs best across all three metrics. Its LPIPS standard deviation reaches 0.235, while CLIP + Attention's is only 0.204, indicating greater perceptual variation in generated images of the same style. The proposed method achieves a Self-FID of 16.7, while ALIGN's is only 19.9, indicating that the internal dis-

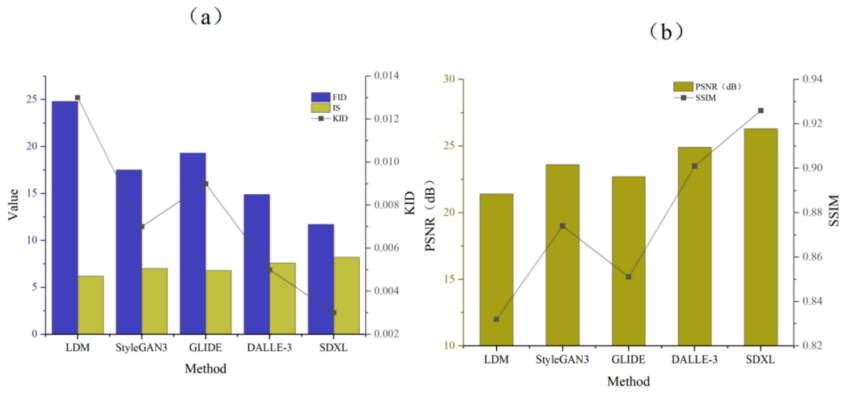

FIGURE 7. Visual quality. Figure 7 (a). FID, IS, and KID results; Figure 7 (b). PSNR and SSIM results

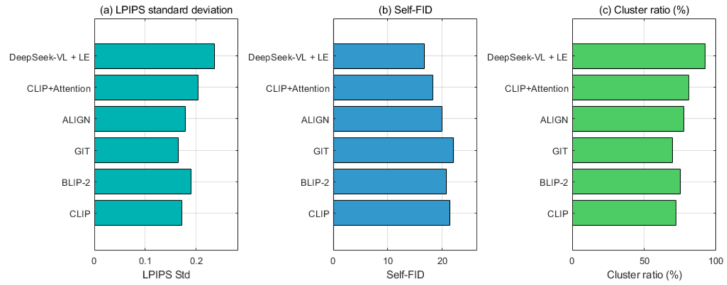

FIGURE 8. Generative diversity evaluation results. Figure 8 (a). LPIPS standard deviation; Figure 8 (b). Self-FID; Figure 8 (c). Proportion of hit (%)

tribution of generated samples is closer to that of the original population and exhibits greater diversity. The most notable achievement is the accuracy rate, with the proposed method achieving 92.5%, far exceeding BLIP-2 (75.1%), GIT (69.8%), and basic CLIP (72.3%). This also represents an 11.3% improvement over CLIP + Attention's 81.2%, demonstrating the strongest coverage of predefined style subclusters.

G. COMPUTATIONAL EFFICIENCY AND RESOURCE CONSUMPTION

The experiment compares the computational efficiency and resource consumption of various methods for the task of generating interior design images. Evaluation metrics include average inference time per image, GPU (Graphics Processing Unit) utilization, model parameter count, and TFLOPs (Tera Floating Point Operations Per Second) to comprehensively measure the model's actual deployment performance and computational burden. The computational efficiency and resource consumption results are shown in Table 3.

TABLE 3. Computational efficiency and resource consumption results

Methods	Average inference time per image (seconds)	GPU utilization (%)	Model parameters (M)	TFLOPs
CLIP (SDXL)	2.7	68.2	890	22.3
BLIP-2 (SDXL)	3.1	72.5	910	23
GIT (SDXL)	2.9	70.1	895	22.8
ALIGN (SDXL)	3	71.3	905	22.9
CLIP + Attention Mechanism (SDXL)	3.6	75.4	950	24.1
DeepSeek-VL + Local Cross-Modal Attention + Example-Driven (SDXL)	4.3	78.7	980	25.5

In Table 3, the average single-image inference time for the basic CLIP (SDXL) solution is 2.7 seconds; BLIP-2 reaches 3.1 seconds; GIT is 2.9 seconds; ALIGN is 3.0 seconds. The attention-based CLIP+Attention mechanism takes 3.6 seconds, while the proposed solution, "DeepSeek-VL + Local Cross-Modal Attention + Example-Driven (SDXL)", takes 4.3 seconds. In terms of GPU utilization, the CLIP solution accounts for 68.2%;

BLIP-2 increases to 72.5%; GIT is 70.1%; ALIGN is 71.3%. CLIP+Attention rises to 75.4%, and the proposed method achieves the highest performance at 78.7%.

In terms of model parameters and TFLOPs, CLIP (SDXL) has 890 million parameters and 22.3 TFLOPs; BLIP-2 and ALIGN increase slightly to 910 million parameters/23.0 TFLOPs and 905 million parameters/22.9 TFLOPs; GIT

has 895 million parameters/22.8 TFLOPs. CLIP+Attention Mechanism (SDXL) has 950 million parameters and 24.1 TFLOPs of computation. This paper's solution integrates DeepSeek-VL, cross-modal attention, and example retrieval, bringing the total number of parameters to 980 million and the computation to 25.5 TFLOPs. Overall, the real-time performance in this paper is not optimal, but it is comparable to CLIP+Attention Mechanism (SDXL), and this paper's solution also achieves the best semantic consistency.

H. MOS RATING

The quality of the images generated by each method is evaluated based on a subjective user survey involving 50 interior design experts and design enthusiasts. A 5-point rating scale (1 = extremely poor, 5 = excellent) is used. Each participant rates the generated images based on six criteria: visual realism, stylistic expression, text matching, composition and layout rationality, overall aesthetics, and color matching and coordination. The average score for each metric is calculated as the Mean Opinion Score (MOS). This study was approved by the Ethics Committee of Hainan Vocational University of Science and Technology, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. All datasets used in this study have undergone strict data anonymization procedures. This process is to protect personal privacy and ensure full compliance with relevant ethical standards and data protection regulations. The results are shown in Figure 9.

In the subjective MOS ratings in Figure 9, the DeepSeek-VL + Local Cross-Modal Attention + Example-Driven (SDXL) method leads significantly in all six dimensions, achieving an average MOS score of 4.65. Visual realism scores 4.6 (up 0.4 from CLIP+Attention's 4.2); style expression scores 4.7 (up 0.6 from ALIGN's 4.1); matching scores 4.8 (up 0.9 from BLIP-2's 3.9); composition and layout rationality scores 4.5 (up 1.0 from CLIP's 3.5); overall aesthetics scores 4.7 (up 1.2 from GIT's 3.5); color matching and coordination scores 4.6 (up 1.2 from CLIP's 3.4). Results demonstrate that this method not only achieves more precise global and local semantic alignment but also better captures the user's textual intent, producing interior renderings with a well-structured layout, harmonious colors, and a high degree of aesthetic appeal.

I. SENSITIVITY ANALYSIS OF THE NUMBER OF RETRIEVED SAMPLES K

To systematically evaluate the impact of the number of retrieved samples k on the generation quality, this experiment fixed other hyperparameters on the IDSSF-64 dataset, adjusting only the number of retrieved samples $k \in \{1,3,5,7,10,15\}$ for the example-driven module. The evaluation used a four-dimensional index system: semantic consistency (CLIP score and DeepSeek-VL similarity), visual quality (FID), style fidelity (ResNet50 classification accuracy), and human perception score (MOS). Table 4 shows the performance comparison under different k values.

TABLE 4. Comparison of Generation Performance with Different Numbers of Retrieved Samples k

k-value	CLIP score	DeepSeek-VL similarity	FID	Style classification accuracy (%)	Mean MOS
1	0.812	0.835	15.3	88.2	3.9
3	0.894	0.917	12.8	93.5	4.3
5	0.936	0.942	11.7	96.0	4.65
7	0.931	0.938	11.9	95.7	4.6
10	0.925	0.931	12.4	94.8	4.5
15	0.907	0.919	13.6	92.3	4.3

Table 4 shows that the value of k has a significant nonlinear impact on the generation quality. When k=1, the system relies excessively on a single example, resulting in insufficient style diversity (style accuracy of only 88.2%) and rigid layout (MOS=3.9). As k increases to 3, semantic consistency and visual quality improve significantly, demonstrating that moderate diversity injection effectively alleviates pattern collapse.

Performance peaks at k=5, where the CLIP score (0.936) and DeepSeek-VL similarity (0.942) are simultaneously optimal, and the FID drops to 11.7, indicating that this setting achieves the best balance between semantic accuracy and visual realism. When k>5, performance shows marginal degradation: the difference between k=7 and k=5 is small, but the noise accumulation effect becomes apparent after k=10, with the FID increasing to 12.4 and the style accuracy decreasing to 94.8%. Notably, the MOS score drops to 4.3 at k=15, with user feedback indicating conflicting decorative elements (such as "mixing Chinese grilles with European carvings") and color inconsistencies in the images. This trend validates the theoretical rationale for k=5—it captures a sufficiently rich variety of style variations while avoiding the semantic dilution risk associated with high k values, and exhibits optimal robustness, especially when dealing with complex scenarios such as mixed styles.

VI. CONCLUSION

Based on the DeepSeek-VL model, this paper constructs a multi-style interior design image generation and semantic consistency enhancement mechanism. DeepSeek-VL is then used for global semantic alignment. A local cross-modal attention module verifies element semantics in a fine-grained manner, and example-driven retrieval is applied to enhance detail representation and diversity. Experimental validation demonstrates that this method significantly outperforms existing technologies in terms of semantic consistency (similarity 0.942), style discrimination (style classification accuracy 96%), and visual quality (FID 11.7, SSIM 0.926), demonstrating excellent multi-style interior design image generation capabilities. However, the model suffers from high computational resource consumption, requires optimization of inference speed, and needs further improvement in generalization to very few sample styles to achieve wider application and practical value.

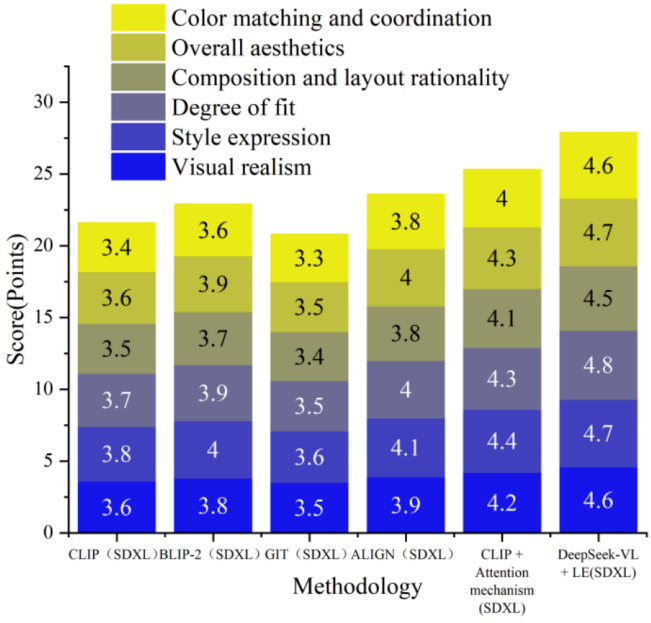

FIGURE 9. MOS rating results

AUTHOR CONTRIBUTIONS

Xiaohui Zhang designed, collected and analyzed the data, and drafted the manuscript. Xiaohui Zhang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Xiaohui Zhang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Hainan Vocational University of Science and Technology, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. All datasets used in this study have undergone strict data anonymization procedures. This process is to protect personal privacy and ensure full compliance with relevant ethical standards and data protection regulations.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] E. Joy and C. Raja, "Digital 3D modeling for preconstruction real-time visualization of home interior design through virtual reality," Construction Innovation, vol. 24, no. 2, pp. 643-653, 2024, doi: 10.1108/CI-10-2020-0174.

[2] H. Lee, S. Je, R. Kim, H. Verma, H. Alavi, and A. Bianchi, "Partitioning open-plan workspaces via augmented reality," Personal and Ubiquitous Computing, vol. 26, no. 3, pp. 609-624, 2022, doi: 10.1007/s00779-019-01306-0.

[3] J. H. Choi and J. S. Lee, "A study of interior style transformation with GAN model," Journal of KIBIM, vol. 12, no. 1, pp. 55-61, 2022, doi: 10.13161/kibim.2022.12.1.055.

[4] Z. Wu, X. Jia, R. Jiang, Y. Ye, H. Qi, and C. Xu, "CSID-GAN: A customized style interior floor plan design framework based on generative adversarial network," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 2353-2364, 2024, doi: 10.1109/TCE.2024.3376956.

[5] J. Chen, Z. Shao, and B. Hu, "Generating interior design from text: A new diffusion model-based method for efficient creative design," Buildings, vol. 13, no. 7, pp. 1861-1878, 2023, doi: 10.3390/buildings13071861.

[6] J. Chen, Z. Shao, X. Zheng, K. Zhang, and J. Yin, "Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation," Scientific Reports, vol. 14, no. 1, pp. 3496-3509, 2024, doi: 10.1038/s41598-024-53318-3.

[7] J. Chen, X. Zheng, Z. Shao, M. Ruan, H. Li, D. Zheng, et al., "Creative interior design matching the indoor structure generated through diffusion model with an improved control network," Frontiers of Architectural Research, vol. 14, no. 3, pp. 614-629, 2025, doi: 10.1016/j.foar.2024.08.003.

[8] H. Cui, L. Liu, H. Zhang, D. Liu, Y. Ma, and Z. Wang, "Detail preserving image generation method based on semantic consistency," Journal of Computer-Aided Design & Computer Graphics, vol. 34, no. 10, pp. 1497-1505, 2022, doi: 10.3724/SP.J.1089.2022.19724.

[9] Y. Ma, L. Liu, H. Zhang, C. Wang, and Z. Wang, "Generative adversarial network based on semantic consistency for text-to-image generation," Applied Intelligence, vol. 53, no. 4, pp. 4703-4716, 2023, doi: 10.1007/s10489-022-03660-8.

[10] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, et al., "Training-free consistent text-to-image generation," ACM Transactions on Graphics (TOG), vol. 43, no. 4, pp. 1-18, 2024, doi: 10.1145/3658157.

[11] Z. Li and R. Zhang, "Direct Distillation: A Novel Approach for Efficient Diffusion Model Inference," Journal of Imaging, vol. 11, no. 2, pp. 66-88, 2025, doi: 10.3390/jimaging11020066.

[12] R. K. Tripti, P. Chethan, S. Arun Vikas, S. S. Monisha, and A. I. Deepseek Open-Source, "International Journal of Trend in Scientific Research and Development," 2025; 9(3):555-560, [Online]. Available: http://eprints.umsida.ac.id/id/eprint/16117.

[13] Z. Deng, W. Ma, Q. L. Han, W. Zhou, X. Zhu, S. Wen, et al., "Exploring

DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions," IEEE/CAA Journal of Automatica Sinica, vol. 12, no. 5, pp. 872-893, 2025, doi: 10.1109/JAS.2025.125498.

[14] J. Wu, J. Chen, J. Wu, W. Shi, X. Wang, and X. He, "Understanding contrastive learning via distributionally robust optimization," Advances in Neural Information Processing Systems, vol. 36, no. 1, pp. 23297-23320, 2023, doi: 10.48550/arXiv.2310.11048.

[15] S. Lee, J. Park, and J. Park, "CrossFormer: Cross-guided attention for multi-modal object detection," Pattern Recognition Letters, vol. 179, no. 1, pp. 144-150, 2024, doi: 10.1016/j.patrec.2024.02.012.

[16] F. Zheng, W. Li, X. Wang, L. Wang, X. Zhang, and H. Zhang, "A cross-attention mechanism based on regional-level semantic features of images for cross-modal text-image retrieval in remote sensing," Applied Sciences, vol. 12, no. 23, pp. 1-19, 2022, doi: 10.3390/app122312221.

[17] H. Tan, B. Yin, K. Xu, H. Wang, X. Liu, and X. Li, "Attention-bridged modal interaction for text-to-image generation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 7, pp. 5400-5413, 2023, doi: 10.1109/TCSVT.2023.3347971.

[18] F. Wang, Y. Su, R. Wang, J. Sun, F. Sun, and H. Li, "Cross-modal and cross-level attention interaction network for salient object detection," IEEE Transactions on Artificial Intelligence, vol. 5, no. 6, pp. 2907-2920, 2023, doi: 10.1109/TAI.2023.3333827.

[19] D. He and C. Xie, "Semantic image segmentation algorithm in a deep learning computer network," Multimedia Systems, vol. 28, no. 6, pp. 2065-2077, 2022, doi: 10.1007/s00530-020-00678-1.

[20] J. Wang, X. Zhang, T. Yan, and A. Tan, "Dpnet: Dual-pyramid semantic segmentation network based on improved deeplabv3 plus," Electronics, vol. 12, no. 14, pp. 3161-3176, 2023, doi: 10.3390/electronics12143161.

[21] D. Lin, Y. X. Peng, J. Meng, and W. S. Zheng, "Cross-modal adaptive dual association for text-to-image person retrieval," IEEE Transactions on Multimedia, vol. 26, no. 1, pp. 6609-6620, 2024, doi: 10.1109/TMM.2024.3355644.

[22] X. Zhao, Y. Tian, K. Huang, B. Zheng, and X. Zhou, "Towards efficient index construction and approximate nearest neighbor search in high-dimensional spaces," Proceedings of the VLDB Endowment, vol. 16, no. 8, pp. 1979-1991, 2023, doi: 10.14778/3594512.3594527.

[23] L. Li, J. Cai, and J. Xu, "A learned index for approximate kNN queries in high-dimensional spaces," Knowledge and Information Systems, vol. 64, no. 12, pp. 3325-3342, 2022, doi: 10.1007/s10115-022-01742-0.

[24] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/s00521-023-08568-z.

[25] I. V. Modoranu, M. Safaryan, G. Malinovsky, E. Kurtić, T. Robert, P. Richtárik, et al., "Microadam: Accurate adaptive optimization with low space overhead and provable convergence," Advances in Neural Information Processing Systems, vol. 37, no. 1, pp. 1-43, 2024, doi: 10.52202/079017-0001.

[26] H. Kabiri, Y. Ghanou, H. Khalifi, and G. Casalino, "AMAdam: Adaptive modifier of Adam method," Knowledge and Information Systems, vol. 66, no. 6, pp. 3427-3458, 2024, doi: 10.1007/s10115-023-02052-9.