IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1405-1417
Journal homepage: http://iieta.org/journals/ts

Fusing Diffusion Models with Structure-Aware Generation Networks for Environment Design Image Synthesis and Realism Enhancement

Mei Bai

Digital Creative Design Institute, Henan Polytechnic, Zhengzhou 450046, China

Corresponding Author Email: baimei2023@163.com

https://doi.org/10.18280/ts.430325

Received: 8 November 2025
Revised: 22 April 2026
Accepted: 30 April 2026
Available online: 30 June 2026

Keywords:
controllable image generation, diffusion models, structure awareness, frequency-domain decoupling, realism enhancement, cross-modal retrieval, environment design image synthesis

ABSTRACT

Artificial intelligence–driven generative technologies are propelling environment design visualization toward greater efficiency and intelligence. However, existing text-to-image diffusion models still suffer from critical limitations when applied to professional environment design image synthesis, impeding the fulfillment of the precision and visual realism required in architectural and landscape design practice. To address these technical bottlenecks, a diffusion-based generation framework that integrates structure awareness and realism enhancement, termed StructDiff-Real, was proposed. Built upon a pre-trained latent diffusion model, a full-chain optimization pipeline was established from three perspectives: conditional encoding, generative sampling, and semantic constraints, tailored to the specific requirements of professional environment design. A multi-scale structural hierarchical encoding mechanism coupled with a progressive feature injection strategy was devised, enabling hierarchical control over global layout, intermediate morphological structures, and fine-grained details, thereby overcoming the scale-adaptation deficiencies inherent in conventional single-modality constraint approaches. Through wavelet transformation, the latent variables were decomposed in the frequency domain, and a differentiated denoising scheduling strategy for structure and texture was constructed. By utilizing a multi-dimensional physically aware discriminator to incorporate physical priors on illumination, shadow, and material properties, structural distortions and texture artifacts in the generated outputs were effectively eliminated. By constructing a domain-specific knowledge base for environment design and employing a cross-modal contrastive learning model, unified embedding and collaborative retrieval-based constraints on semantic and structural features were realized, significantly mitigating semantic drift caused by abstract design inputs. Comprehensive comparative experiments on both professional datasets and public benchmark datasets demonstrated that the proposed method substantially outperformed current state-of-the-art approaches in terms of structural fidelity, visual realism, and cross-modal semantic alignment. This research provides technical support for rapid visual prototyping and intelligent multi-scheme iteration in environment design while offering a novel paradigm for controllable, high-fidelity image generation in professional design scenarios.

1. INTRODUCTION

With the widespread adoption of digital twin technologies and building information systems, the environment design industry has been progressively transitioning away from traditional manual drafting and offline rendering workflows, moving toward digital, intelligent, and rapidly iterative paradigms [1-3]. In the engineering practice of contemporary architectural landscape and site environment design, the efficiency of visual prototyping for design schemes constitutes a core factor that constrains iterative optimization. Conventional design visualization [4-6], which relies on the complete pipeline of three-dimensional modeling, material tuning, and light baking, involves cumbersome procedures and protracted cycles, rendering multi-scheme comparative iteration extremely costly and ill-suited to the current industry demand for high-frequency and high-efficiency design creation. In recent years, text-driven diffusion models, by virtue of their powerful cross-modal generation capabilities, have become a cornerstone technology in the field of image generation, demonstrating exceptional performance in general-purpose visual image creation. However, pre-trained diffusion models for general purposes exhibit poor adaptability when applied to environment design—a professional domain characterized by high precision and strong structural regularity [7-9]. These models lack dedicated constraint mechanisms tailored to architectural spatial layout, geometric structural forms, and the physical properties of lighting and materials. Consequently, the generated outputs commonly suffer from structural irregularities, insufficient visual realism, and deviation from design intentions. In response to these industry pain points and technical

deficiencies, the controllable and high-fidelity generation of professional environment design images is addressed, and a dedicated diffusion-based generation and enhancement framework is constructed. This investigation not only contributes to advancing the theoretical framework of controllable image generation in professional scenarios and filling the technical gaps in multi-scale structural precision control and physical realism modeling, but also provides actionable technical support for intelligent environment design creation and rapid visual prototyping of design schemes, thereby bearing significant academic theoretical value and engineering application potential.

Although substantial progress has been achieved in conditional image generation and visual realism enhancement techniques, three critical technical bottlenecks persist when these methods are deployed for professional environment design image synthesis, severely compromising the practical usability of the generated outcomes. The first bottleneck pertains to scale-adaptation deficiencies in structural control. Existing conditional constraint generation methods predominantly adopt a unimodal, unified feature injection paradigm, without accounting for the inherently hierarchical structural characteristics of environment design images [10-12]. These methods are incapable of differentiating generation priorities and constraint weights among global spatial layout, intermediate architectural massing, and local detailed elements, nor are they able to dynamically modulate constraint intensity in accordance with the feature evolution patterns observed across distinct denoising stages of diffusion models. Consequently, either global compositional imbalance or local structural omissions are induced, precluding simultaneous assurance of generation fidelity across multiple structural scales. The second bottleneck concerns generation defects arising from frequency-domain feature coupling. The denoising pathways of existing diffusion models perform synchronous modeling of low-frequency structural information and high-frequency textural information [13, 14], wherein the two categories of features interfere with and compete against each other during optimization, readily leading to structural distortions, edge dislocations, and texture artifact superimposition in the generated images. Concurrently, effective physical prior constraints are absent from current generation methodologies, with core physical attributes such as illumination consistency [15, 16], shadow occlusion logic, and material reflectance laws [17, 18] failing to be modeled. As a result, the visual presentation of generated images deviates from authentic physical rules, rendering it difficult to satisfy the realism requirements of professional design. The third bottleneck lies in cross-modal semantic alignment deviations. Designers' commonly employed hand-drawn sketches and brief textual descriptions are characterized by semantic sparsity and informational ambiguity. Existing models have not incorporated domain-specific prior knowledge from environment design, nor do they possess cross-modal alignment constraint mechanisms bridging structure and semantics. As a consequence, the design intentions embedded in abstract inputs cannot be precisely captured, and the generated results are prone to semantic drift, disorganized functional layouts, and low stylistic matching [19, 20], ultimately failing to serve practical design workflows.

In response to the aforementioned industry pain points and technical limitations, an integrated structure-aware realism-enhanced diffusion generation framework is proposed, yielding multidimensional innovative contributions. A hierarchical structural encoding and progressive feature injection mechanism is constructed, enabling refined controllable generation of multi-scale structures in environment design. A frequency-domain decoupled sampling strategy coupled with a physically aware discriminator constraint method is introduced, effectively enhancing the physical realism of the generated images. A cross-modal retrieval-augmented alignment system is established to address the semantic drift problem under abstract design inputs. Furthermore, a professional environment design test benchmark is built, and the effectiveness and superiority of the proposed methodology are systematically validated through multiple sets of quantitative and qualitative experiments.

A latent diffusion model is adopted as the foundational generative backbone, and an integrated structure-aware realism-enhanced diffusion generation framework is constructed to address the prevalent issues of insufficient structural control precision, lack of physical realism, and design semantic drift in environment design image synthesis tasks. The proposed framework relies on a complete technical pipeline encompassing conditional encoding, hierarchical feature injection, frequency-domain decoupled generation, and cross-modal alignment constraints, with systematic optimizations performed across three dimensions: model input, generative inference, and global constraints. An end-to-end generation system from raw design inputs to high-quality professional design image outputs is thereby established, which is capable of precisely accommodating the dual professional requirements of structural precision and visual realism in environment design scenarios.

2. STRUCTDIFF-REAL: STRUCTURE-AWARE REALISM-ENHANCED DIFFUSION FRAMEWORK

2.1 Framework overview

A latent diffusion model is adopted as the foundational generative backbone, and an integrated structure-aware realism-enhanced diffusion generation framework is constructed to address the prevalent issues of insufficient structural control precision, lack of physical realism, and design semantic drift in environment design image synthesis tasks. A schematic illustration of the overall architecture and full-chain optimization of StructDiff-Real is presented in Figure 1. The proposed framework relies on a complete technical pipeline encompassing conditional encoding, hierarchical feature injection, frequency-domain decoupled generation, and cross-modal alignment constraints, with systematic optimizations performed across three dimensions: model input, generative inference, and global constraints. An end-to-end generation system from raw design inputs to high-quality professional design image outputs is thereby established, which is capable of precisely accommodating the dual professional requirements of structural precision and visual realism in environment design scenarios.

The core advantage of this framework lies in the deep synergistic coupling among its functional modules, rather than the mere superposition of independent functionalities. On the input side, the multi-level structural disassembly and encoding mechanism is designed to provide stable and reliable multi-scale structural anchor points for the generation process, thereby furnishing foundational geometric constraints for the subsequent frequency-domain decoupling optimization. On

the generation side, the frequency-domain decoupled sampling and physical realism enhancement strategy is capable of optimizing textural details and physical lighting performance while preserving structural integrity, thereby eliminating generation defects arising from feature coupling. On the constraint side, the cross-modal retrieval-augmented mechanism, grounded in domain-specific professional priors, simultaneously calibrates structural morphology and design semantics, compensating for the incompleteness of abstract input information. Through feature propagation and joint loss constraints, the individual modules are integrated into an organic whole, fundamentally overcoming the adaptation deficiencies of general-purpose diffusion models in professional environment design generation scenarios.

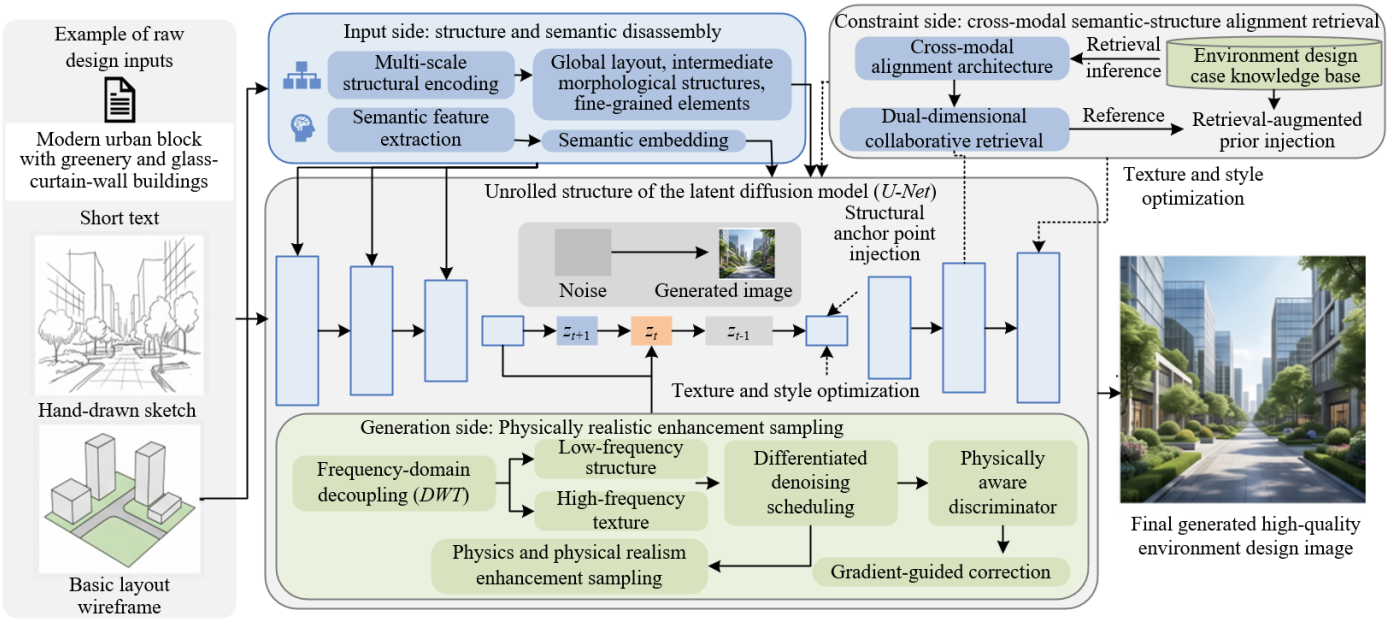

Figure 1. Overall architecture and full-chain optimization schematic of StructDiff-Real

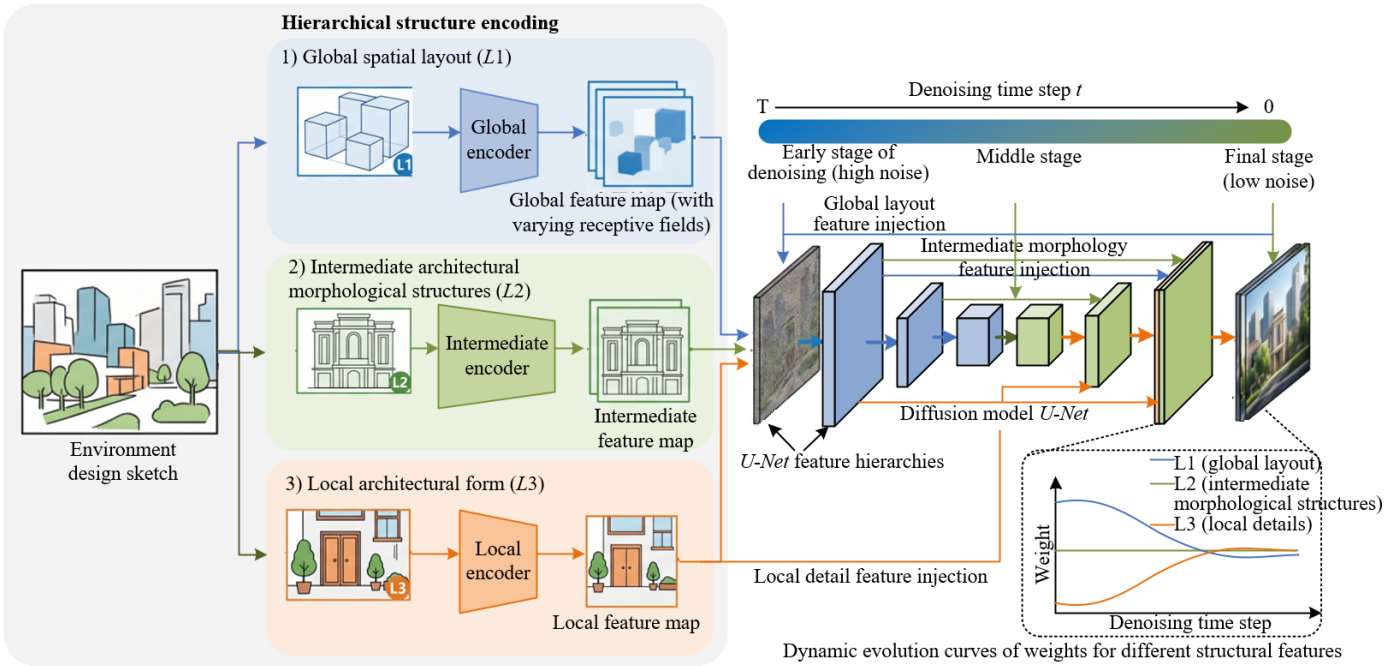

Figure 2. Multi-scale structural hierarchical encoding and progressive feature injection mechanism

To address the scale-adaptation limitations inherent in conventional unimodal feature injection approaches, a multi-scale structural hierarchical encoding and progressive feature injection mechanism is designed at the input side of the proposed framework, with the specific workflow illustrated in Figure 2. Through this mechanism, the underlying design sketch is decomposed along the spatial scale into three semantic hierarchies: global layout, intermediate morphological structures, and local details, which correspond respectively to macroscopic site massing, building facade setback profiles, and microscopic elements such as windows, doors, and landscape accessories. Three parallel branches are employed within the network to process information from each hierarchy independently, generating dedicated feature maps adapted to different receptive fields. In conjunction with the intrinsic denoising evolution patterns of the diffusion

model, the feature injection process is governed by a time-step-dependent dynamic weight scheduling strategy. During the early denoising stage under high-noise conditions, the highest weight is assigned to global layout features to anchor the macroscopic composition. As the iterative process proceeds, the control emphasis is smoothly transferred to intermediate morphological features. In the final low-noise denoising stage, the injection intensity of local detailed features reaches its peak, serving to refine architectural components and environmental details. Through this time-varying fusion strategy—characterized by decay and enhancement over time steps—cross-scale feature interference is effectively decoupled, and rigorous hierarchical precision control is achieved along the spatial dimension.

2.2 Structure-texture frequency-domain decoupled sampling strategy for realism enhancement

The denoising iterative process of existing latent diffusion models performs joint modeling of structural and textural features, with the optimization objectives of these two feature categories exhibiting inherent conflicts. Structural features prioritize global consistency in spatial layout and geometric topology, whereas textural features focus on local richness in surface details, illumination variations, and material expressions. A unified denoising constraint inevitably induces mutual interference between these features, rendering it difficult to simultaneously satisfy the dual professional requirements of structural precision and visual realism in environment design images. To address this problem, a two-dimensional discrete wavelet transform is introduced into the latent space of the diffusion sampling process, enabling complete decoupling of structural and textural features. For the latent variable z_t at any given sampling time step, frequency-domain decomposition is performed, yielding four orthogonal frequency components:

$$\{z_t^{LL}, z_t^{LH}, z_t^{HL}, z_t^{HH}\} = DWT(z_t) \tag{1}$$

where, z_t^{LL} denotes the low-frequency approximation component, which carries global spatial layout and geometric structural information and constitutes the core feature determining the morphological integrity of the design scheme; and z_t^{LH}, z_t^{HL}, and z_t^{HH} are high-frequency detail components, corresponding respectively to edge textures and illumination details in the horizontal, vertical, and diagonal directions. Accordingly, the latent variable can be reconstructed into two independent representations—low-frequency structural components and high-frequency textural components—thereby enabling separate optimization of features with distinct attributes. Compared to decomposition in pixel space, the latent-space approach operates in a lower-dimensional space with reduced redundant information, effectively lowering computational overhead while aligning with the native latent-space modeling mechanism of diffusion models and circumventing feature degradation caused by repeated encoding and decoding operations.

Based on the frequency-domain decoupled feature representation system, a time-varying adaptive guidance weight mechanism is established, with differentiated denoising optimization strategies designed for structural and textural components, thereby constructing a progressive generation iterative logic. The overall denoising optimization formulation can be defined as:

$$\begin{aligned}\epsilon_\theta(z_t,t,C) = \epsilon_\theta(z_t,t) + \gamma_{struct}(t)\cdot\nabla_{z_t^{LL}}\log p(z_t^{LL}|C) \\ + \gamma_{texture}(t)\cdot\nabla_{z_t^{texture}}\log p(z_t^{texture}|C)\end{aligned} \tag{2}$$

where, ϵ_θ denotes the noise residual predicted by the model, C denotes the input structural constraint condition, and $\gamma_{struct}(t)$ and $\gamma_{texture}(t)$ denote the dynamic guidance weights for low-frequency structural and high-frequency textural components, respectively. The weight parameters exhibit a linear monotonic variation pattern with respect to the diffusion time step t. During the early stage of denoising iteration, global structural composition is prioritized, with a high structural weight value employed to reinforce geometric constraints and suppress interference from high-frequency noise on contour formation. In the later iterative stage, the structural contours have largely stabilized, and the structural constraint weight is gradually reduced while the textural guidance weight is concurrently increased, enabling refined optimization of material details and lighting performance. The optimal scheduling intervals for the two weight sets are determined through a grid search method, achieving a dynamic balance between structural stability and textural richness and fundamentally eliminating generation defects arising from frequency-domain feature coupling at the generative mechanism level. The structure-texture frequency-domain decoupled sampling and physically aware discriminative constraint strategy are illustrated in Figure 3.

To further enhance the physical plausibility of texture generation in alignment with optical and material principles governing real-world environmental scenes, a multi-dimensional physically aware discriminator is constructed to enable refined evaluative constraints on illumination consistency, shadow compliance, and material authenticity. The discriminator takes the generated image as input and outputs a four-dimensional quantitative scoring vector, thereby achieving comprehensive characterization of the physical attributes of the image:

$$s = \Psi(x) = [s_{light}, s_{shadow}, s_{material}, s_{global}]^\top \tag{3}$$

where, s_{light} characterizes the uniformity of global illumination direction, computed from the statistical variance of highlight features across multiple regions of the image; s_{shadow} quantifies the degree of matching between shadow morphology and occluding object scales, reflecting the physical compliance of shadow generation; $s_{material}$ evaluates the naturalness of surface reflectance properties based on statistical features derived from the bidirectional reflectance distribution function; and s_{global} comprehensively represents the overall visual texture quality of the image. The discriminator is trained through contrastive learning on real environment design images and model-synthesized images, with a gradient penalty mechanism incorporated to ensure training convergence stability. The overall loss function is formulated as:

$$\begin{aligned}L_\Psi = E_{x_{real}}[\|\Psi(x_{real})-1\|_2^2] \\ + E_{x_{fake}}[\|\Psi(x_{fake})-0\|_2^2] + \beta\cdot\|\nabla_x\Psi(x)\|_2^2\end{aligned} \tag{4}$$

where, x_{real} and x_{fake} denote real samples and generated samples, respectively; 1 and 40 denote the ideal real-score and fake-score vectors, respectively; and β denotes the gradient

penalty coefficient, which is employed to constrain the magnitude of parameter updates, thereby preventing training gradient explosion and mode collapse, and ensuring stable convergence of the discriminator's evaluation accuracy across various physical attributes.

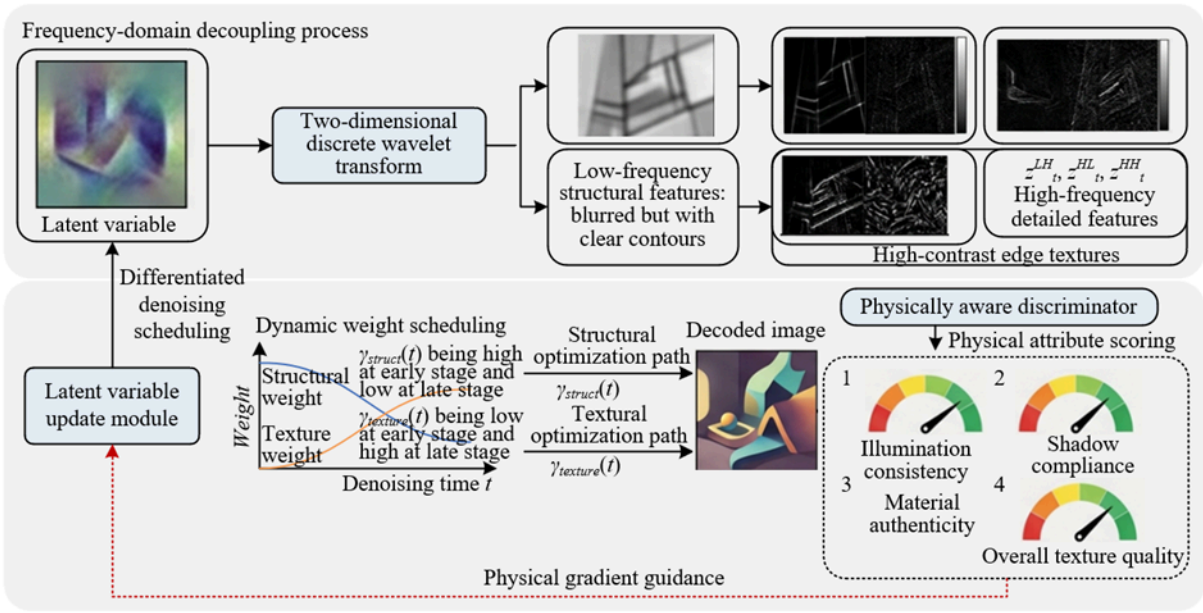

Figure 3. Structure-texture frequency-domain decoupled sampling and physically aware discriminative constraint strategy

To effectively integrate physical prior knowledge into the diffusion sampling process, a gradient guidance mechanism is constructed based on the discriminator's realism scores, enabling refined correction of the latent variable denoising iteration direction. The optimized latent variable update rule is formulated as:

$$z_{t-1}=z_{t-1}^{base}+\eta\cdot\nabla_{z_t}\|\Psi(D(z_t))-1\|_2^2 \tag{5}$$

where, z_{t-1}^{base} denotes the latent variable output by the standard diffusion sampling pipeline; D denotes the pre-trained variational autoencoder decoder, which is responsible for mapping the latent variable to pixel space for realism evaluation; and η denotes the realism guidance intensity coefficient. Considering that the overall image structure has not yet been fully formed during the early iterative stages, and premature introduction of physical constraints would interfere with the generation accuracy of spatial layout, the physical gradient guidance is activated only during the final thirty percent of the denoising process iterations. Through this temporal constraint strategy, precise optimization of physical properties—including textures, illumination, and materials—during the later stages is achieved while ensuring stable generation of multi-scale structures, thereby realizing synergistic optimization of structural fidelity and physical realism in environment design images, and comprehensively enhancing the professional applicability of the generated results.

2.3 Retrieval-augmented mechanism with cross-modal semantic-structure alignment

In environment design image generation scenarios, the input information constituted by hand-drawn sketches and brief textual descriptions is generally characterized by sparsity and ambiguity. Relying solely on basic conditional constraints is insufficient for precisely anchoring the designer's creative intentions, readily leading to semantic drift and disorganized design logic in the generated outcomes. To compensate for the informational deficiencies of abstract input conditions, the introduction of domain-specific professional prior knowledge constitutes an effective approach for enhancing the compliance of generated content. To this end, a standardized and structured environment design case knowledge base is constructed, providing reliable structural priors and stylistic textural priors for the diffusion generation process. The overall standardized definition of the knowledge base is formulated as:

$$K=\{(I_i,S_i,T_i)\}_{i=1}^{N} \tag{6}$$

where, N denotes the total number of samples in the knowledge base; I_i denotes a high-quality real environment design image covering mainstream design scenarios such as building facades, landscape sites, and streetscape interfaces; S_i denotes the corresponding three-level hierarchical structural annotations, strictly adhering to the previously established structural partitioning rules of global layout, intermediate morphological structures, and local details, thereby ensuring consistency within the structural system; and T_i denotes a standardized semantic tag system encompassing three core categories of semantic information: scene style, functional attributes, and material characteristics. All samples and their annotations have undergone dual verification by professional designers, with erroneous annotations and low-quality samples eliminated, thereby ensuring the precision and professionalism of the prior information contained in the knowledge base.

To overcome the spatial heterogeneity barriers between structural and semantic modalities, and to achieve collaborative retrieval and unified constraints across the two types of information, a dual-encoder cross-modal feature alignment architecture is established, enabling the embedding mapping of heterogeneous information into a common

dimensional space. Through two independent encoders, the mixed information from structural conditions and semantic sketches is respectively parsed, and the two categories of features are mapped into a shared embedding space of uniform dimensionality. The feature encoding process can be formulated as:

$$h_{struct}=E_{struct}(C),\ h_{sem}=E_{sem}(T,Sketch) \tag{7}$$

where, E_{struct} denotes the structural encoder, responsible for transforming the hierarchical structural conditions C into high-dimensional structural features h_{struct}; and E_{sem} denotes the cross-modal semantic encoder, which fuses textual semantics T with visual features from hand-drawn sketches to output unified semantic features h_{sem}. The output dimensions of both feature types are unified to d. On this basis, information noise-contrastive estimation contrastive learning is adopted to achieve precise cross-modal feature alignment, whereby feature discrepancies between matched modalities are reduced through distance constraints on positive and negative sample pairs. The loss function is defined as:

$$L_{align}=-\log\frac{exp(sim(h_{struct},h_{sem}^{+})/\tau)}{\sum_{j=1}^{B}exp\,(sim(h_{struct},h_{sem}^{j})/\tau)} \tag{8}$$

where, $sim(\,,\,)$ denotes the cosine similarity function; τ denotes the temperature coefficient, which is used to smooth the feature distribution and modulate gradient sensitivity; B denotes the training batch size; h_{sem}^{+} denotes the positive semantic feature that precisely matches the current structural features; and h_{sem}^{j} denotes the negative sample features within the same batch. Through iterative optimization of this loss, the structural and semantic modalities can be brought into deep spatial alignment, thereby providing foundational support for subsequent collaborative retrieval.

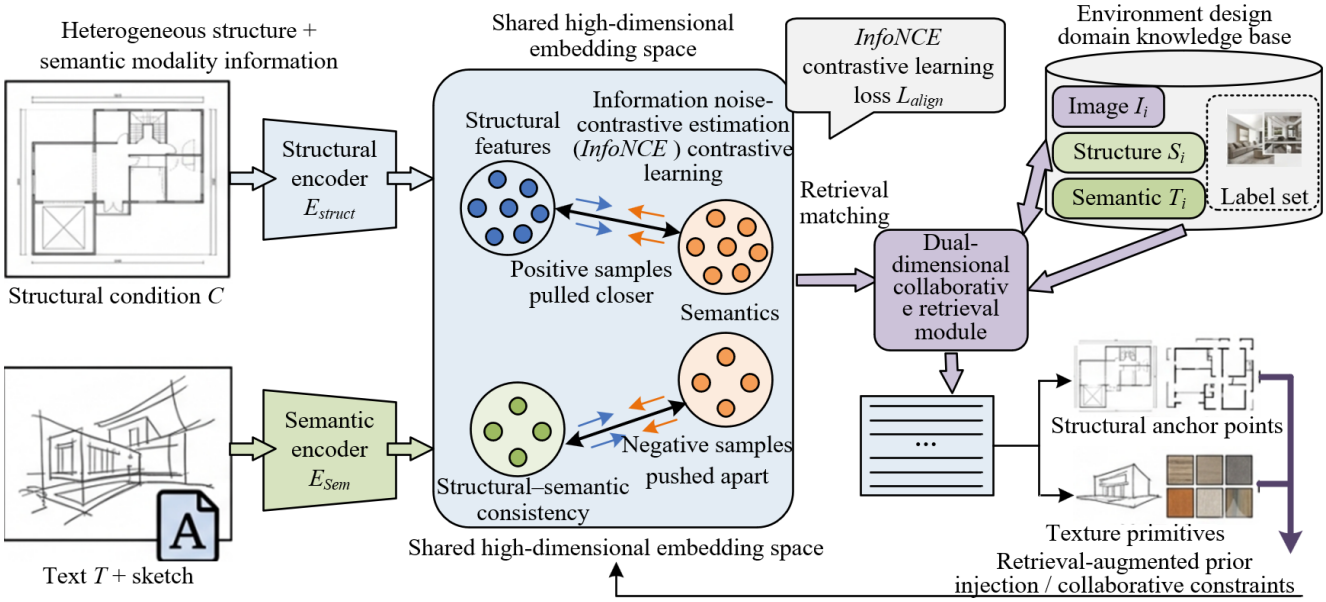

Figure 4. Dual-encoder cross-modal feature alignment and domain knowledge collaborative retrieval architecture

Based on the aligned shared embedding space, a dual-dimensional collaborative retrieval mechanism integrating structure and semantics is established, overcoming the one-sided deficiencies inherent in single-modality retrieval and achieving comprehensive and precise matching of reference samples. The dual-encoder cross-modal feature alignment and domain knowledge collaborative retrieval architecture are illustrated in Figure 4. During the inference stage, the similarity between structural features and semantic features is computed synchronously, and the comprehensive matching score is obtained through equal-weight fusion, from which the optimal set of reference samples best suited to the current design requirements is selected. The retrieval rule is formulated as:

$$R=\underset{(I_i,S_i,T_i)\in K}{top\text{-}K}\left[\begin{matrix}sim(E_{struct}(C),E_{struct}(S_i))\\+sim(E_{sem}(T),E_{sem}(T_i))\end{matrix}\right] \tag{9}$$

Through this mechanism, both spatial structural rationality and semantic stylistic consistency of the generated outcomes are simultaneously accounted for, effectively avoiding the style mismatch problems caused by structure-only retrieval as well as the layout disorganization issues induced by semantics-only retrieval, thereby providing high-quality and highly relevant domain prior samples for the generation process.

The retrieved reference samples are integrated into the diffusion generation pipeline through a dual-pathway approach, enabling full-chain optimization from both structural constraints and stylistic textural dimensions, thereby establishing a consistency constraint system that synergistically combines semantic and structural alignment. At the structural level, the hierarchical structural annotations of the reference samples are leveraged to supplement geometric anchor points, reinforcing the professional regularity of the generated layout. At the textural level, color statistics and texture primitive features are extracted from the reference images, and the high-frequency texture denoising process is optimized through adaptive instance normalization, thereby enhancing the domain-specific adaptability of material and lighting representations. A joint loss function that integrates semantic matching and structural alignment is constructed to perform global optimization on the generated results:

$$L_{sem\text{-}struct}=CE(Sem(D(z_t)),T) +\lambda_{ref}\cdot\sum_{(I_r,S_r)\in R} dist\,(Struct(D(z_t)),S_r) \quad (10)$$

where, CE denotes the cross-entropy loss, employed to quantify the degree of semantic matching between the generated image and the design text; D denotes the variational autoencoder decoder, which performs the mapping from latent variables to pixel space; $Sem(\,)$ and $Struct(\,)$ denote the semantic classification operator and structural feature extraction operator, respectively; $dist(\,,\,)$ denotes the feature cosine distance function; and λ_{ref} denotes the reference structural constraint weight. Through this loss, both semantic drift and structural distortion issues are simultaneously corrected, fully leveraging the prior constraint capacity of the domain knowledge base and significantly enhancing the stability and professional quality of generated outcomes in complex design scenarios.

2.4 Unified training objective and inference pipeline

The multiple optimization objectives—encompassing diffusion denoising, structural constraints, physical realism discrimination, cross-modal alignment, and retrieval-based semantic constraints—are integrated into a multi-task joint optimization loss function, enabling coordinated and balanced learning across all functional modules. The overall unified loss function is formulated as:

$$L_{total}=L_{diff}+\lambda_1 L_{struct}+\lambda_2 L_{\Psi}+\lambda_3 L_{align}+\lambda_4 L_{sem\text{-}struct} \quad (11)$$

where, L_{diff} denotes the foundational denoising loss of the latent diffusion model, which preserves the native image generation capacity of the model; L_{struct} denotes the multi-scale structural consistency loss, which constrains the spatial layout and geometric topological precision of the generated images; L_{Ψ} denotes the physically aware discriminative loss, responsible for optimizing the physical plausibility of illumination, shadows, and materials; L_{align} denotes the cross-modal feature alignment loss, which achieves feature-space unification between structural and semantic modalities; and $L_{sem\text{-}struct}$ denotes the semantic-structural collaborative constraint loss, which rectifies semantic drift in the generated outcomes. The coefficients λ_1, λ_2, λ_3, and λ_4 denote tunable balancing weights for each sub-loss term. The optimal weight configuration is determined through a grid search strategy on the validation set, effectively harmonizing the optimization gradients across different tasks, preventing any single task from dominating model training, and ensuring balanced improvement across all generative performance dimensions.

To address the issues of gradient competition, parameter oscillation, and convergence instability arising from multi-task joint training, a progressive three-stage training strategy is designed, in which parameter convergence and collaborative fine-tuning are performed module by module according to functional priority. During the initial training stage, all parameters of the pre-trained diffusion backbone are frozen, and only the multi-branch structure-aware control network and hierarchical adaptation weights are iteratively optimized. Relying on the foundational diffusion loss and structural constraint loss, prioritized convergence of structural control capability is achieved, ensuring that the model possesses stable multi-scale geometric constraint capacity. In the second stage, the backbone network and structural control module parameters are fixed, and the physically aware discriminator is independently trained until the loss stabilizes. The discriminator gradient guidance mechanism is then embedded into the sampling pipeline for fine-tuning, establishing the optimization capability for image physical realism. In the third stage, the parameters of the previously matured modules are locked, and cross-modal contrastive learning training of the dual encoders is performed, achieving precise alignment between structural and semantic features. Finally, a minor end-to-end fine-tuning of the global model parameters is conducted, enabling deep coupling among the three major modules—structural control, frequency-domain realism enhancement, and retrieval alignment—to accommodate the generation requirements of complex environment design images.

During the model inference stage, a complete and orderly high-fidelity image generation pipeline is formed through the collaborative interaction of the various modules. The input sketch, layout wireframe, and textual description are first preprocessed, from which standardized three-level hierarchical structural conditions and semantic representations are parsed. Feature embedding is then performed through the dual encoders, and the optimal reference samples are retrieved from the knowledge base. Throughout the diffusion sampling iterations, multi-scale structural features are integrated into the U-Net feature hierarchies according to the progressive injection rules, while frequency-domain decoupling optimization is simultaneously applied to the latent variables at each step, with low-frequency structural shaping and high-frequency textural refinement respectively accomplished through dynamic weight scheduling. In the later sampling stages, physically aware gradient constraints and retrieval-based prior alignment constraints are introduced to correct the denoising direction of the latent variables, further enhancing image physical plausibility and design semantic matching. Finally, latent variable decoding is performed through the pre-trained variational autoencoder, outputting environment design images that simultaneously exhibit structural precision, physical realism, and design semantic consistency.

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental setup

To comprehensively validate the effectiveness and generalization capability of the proposed method in professional environment design scenarios, an experimental scheme combining a self-constructed professional dataset and publicly available general-purpose datasets was adopted. A high-quality environment design dataset, designated EnvDesign-HQ, was constructed, comprising a total of 8,000 high-resolution environment design images covering mainstream application scenarios such as building facade design, site master planning, outdoor landscape construction, and interior-exterior spatial design. The dataset was simultaneously equipped with a three-level structured annotation system—encompassing global layout, intermediate massing, and local details—alongside standardized semantic description texts, enabling precise alignment with the model's requirements for multi-scale structural constraints and cross-modal alignment training. The dataset was randomly partitioned into training and test sets at a 4:1 ratio, with 6,400

samples used for model training and parameter fitting, and 1,600 samples reserved for performance testing and metric evaluation. To further verify the cross-scenario generalization performance of the model, the ADE20K architectural subset and the Places365 landscape subset were selected as general-purpose validation datasets, enabling dual-performance verification across both professional and general visual scenarios.

A multi-dimensional and hierarchical quantitative evaluation system was constructed to objectively assess generated image quality from three core dimensions: structural fidelity, visual realism, and semantic alignment, with professional subjective evaluation additionally introduced to achieve consistency between quantitative metrics and industry cognition. Structural fidelity was employed to measure the precision of geometric structure and spatial layout in the generated images, specifically encompassing three metrics: edge intersection over union, mean intersection over union for semantic segmentation, and layout boundary matching accuracy. Visual realism was employed to characterize the physical plausibility and visual quality of illumination, textures, and materials in the images, with quantitative evaluation performed using the Fréchet inception distance, kernel inception distance, and inception score. Semantic alignment was employed to measure the degree of matching between the generated outcomes and the input design intentions, with quantitative analysis conducted through contrastive language–image pre-training text-image similarity and semantic label classification accuracy. For the subjective evaluation, 15 professional environment design practitioners with years of industry experience were invited to conduct blind evaluation experiments, scoring the generated results on a 100-point scale across three dimensions—structural accuracy, material realism, and design conformance—to objectively reflect the professional applicability of the generated outcomes.

To comprehensively verify the superiority of the proposed method, seven state-of-the-art approaches currently prevalent in the field were selected as comparative baselines, covering three major technical directions: general-purpose diffusion generation, structured conditional constraints, and visual realism enhancement. These specifically included representative models such as Stable Diffusion, ControlNet, Text-to-Image Adapter, LayoutDiffusion, and RealDiffusion. All comparative methods were implemented using their official open-source code and default hyperparameter configurations, with training and testing conducted under unified hardware and dataset environments to ensure the fairness and validity of the comparative experiments. The proposed model was built upon the PyTorch deep learning framework, with the training batch size set to 16, the initial learning rate set to 1×10^{-4}, the total number of training epochs set to 120, and the number of diffusion inference sampling steps set to 50. All experiments were uniformly deployed on a single NVIDIA RTX 4090 graphics processing unit device, with hyperparameter combinations traversed through grid search to determine the optimal values for each loss weight and guidance coefficient. All experimental procedures were strictly standardized to ensure reproducibility of the results.

3.2 Overall performance comparison experiments

Comprehensive performance comparisons between the proposed StructDiff-Real framework and all baseline models were conducted on the self-constructed EnvDesign-HQ dataset and the public ADE20K dataset, to thoroughly validate the integrated advantages of the model in structural control, visual generation, and semantic matching. Among the metrics, higher values for edge intersection over union, mean intersection over union, layout matching accuracy, inception score, contrastive language–image pre-training similarity, and semantic accuracy indicate superior performance, whereas lower values for Fréchet inception distance and kernel inception distance indicate better image realism and visual quality.

From the overall quantitative results presented in Table 1, it can be observed that the proposed framework achieves significantly superior performance over all existing state-of-the-art baseline models across every evaluation metric, attaining simultaneous optimal performance in the three dimensions of structure, realism, and semantics. Compared with LayoutDiffusion, a model focused on structural constraints, the proposed method achieves improvements of 12.7% in edge intersection over union and 12.2% in mean intersection over union, along with an increase of nearly 9 percentage points in layout matching accuracy, demonstrating that the hierarchical progressive structure injection mechanism effectively overcomes the inherent limitations of traditional methods in accommodating multi-scale structures. Compared with RealDiffusion, a model centered on realism optimization, the proposed method substantially reduces Fréchet inception distance and kernel inception distance while significantly improving the inception score, fully validating the optimization effects of frequency-domain decoupled sampling and physically aware constraints on texture quality and physical plausibility. In the semantic alignment dimension, the model achieves contrastive language–image pre-training text-image similarity and semantic classification accuracy substantially exceeding those of all comparative models, effectively resolving the semantic drift problem under abstract design inputs. From the qualitative visualization results, it is evident that images generated by baseline models commonly suffer from structural distortions, element disorganization, lighting conflicts, and material artifacts. In contrast, the environment design images generated by the proposed method exhibit well-regulated spatial layouts, precise geometric structures, and lighting and material properties conforming to physical laws, closely matching the creative intentions of designers and fully meeting the visualization standards of professional environment design.

To verify whether the proposed method can simultaneously achieve structural controllability, texture enhancement, and realism correction in the actual environment design image synthesis pipeline, a visualization analysis of the complete generation process was performed using the corresponding implementation effect renderings, as shown in Figure 5. As can be observed from Figure 5, the method first performs hierarchical parsing of the architectural massing, courtyard boundaries, paving areas, and landscape details from the original design sketch, enabling the abstract input to be transformed into structural constraint information that can be invoked by the diffusion generation process. In the initial diffusion generation stage, the image already possesses basic scene semantics, yet issues such as building edge deviation, local texture artifacts, and inconsistent lighting relationships remain, indicating that reliance on the diffusion model alone is insufficient to meet the professional requirements of environment design images in terms of geometric precision

and physical realism. After frequency-domain decoupling and texture enhancement, the low-frequency structure is stably preserved, while high-frequency material details are notably refined, with improved visual continuity observed in paving, water surfaces, vegetation, and building surface textures. Following the further introduction of physical realism correction, the lighting direction, shadow boundaries, and material reflectance relationships become more consistent, elevating the generated images from visual plausibility to physical credibility. The final results exhibit good structural clarity, textural refinement, and semantic consistency in localized regions such as building facades, stone and glass materials, vegetated water features, and paving transitions, demonstrating that the proposed method—integrating diffusion models with structure-aware generation networks—effectively addresses the structural drift, material distortion, and insufficient realism commonly encountered by traditional generation models in professional environment design scenarios.

Table 1. Quantitative overall performance comparison results of different methods

Method	Edge Intersection over Union	Mean Intersection over Union	Layout Matching Accuracy (%)	Fréchet Inception Distance	Kernel Inception Distance	Inception Score	Contrastive Language–Image Pre-Training Similarity	Semantic Accuracy (%)
Stable Diffusion	0.612	0.587	72.36	23.15	0.082	7.26	0.713	75.21
ControlNet	0.725	0.693	81.52	18.62	0.061	8.13	0.765	80.36
Text-to-Image Adapter	0.701	0.671	79.85	19.45	0.065	7.89	0.752	78.92
LayoutDiffusion	0.742	0.715	83.21	17.93	0.058	8.25	0.771	81.25
RealDiffusion	0.685	0.662	78.63	15.26	0.049	8.62	0.748	79.63
Proposed method	0.836	0.802	92.15	10.32	0.026	9.58	0.856	90.12

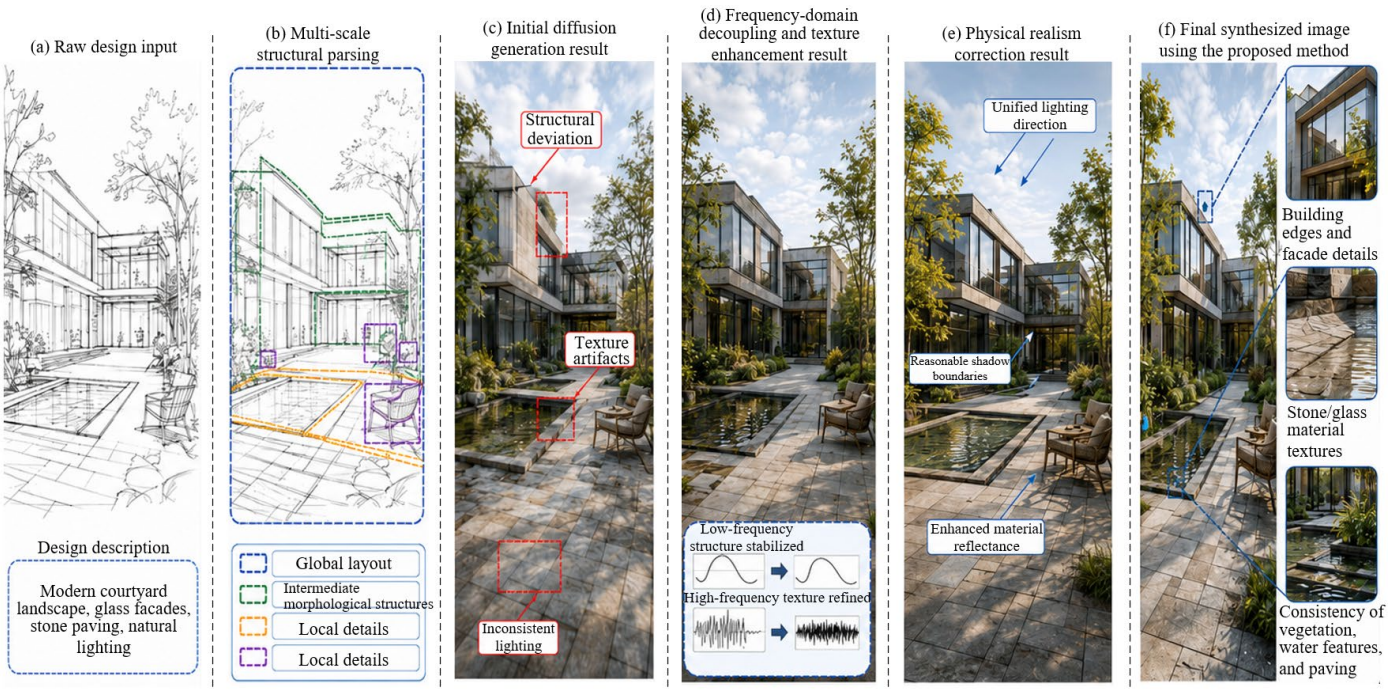

Figure 5. Implementation effect renderings of StructDiff-Real for environment design image synthesis and realism enhancement

3.3 Module ablation experiments

To verify the individual contributions and synergistic optimization effects of the three core innovative modules, a stepwise ablation experimental framework was established. Stable Diffusion combined with a single-branch ControlNet was adopted as the baseline model, upon which the multi-level structural progressive injection module, frequency-domain decoupled sampling module, physically aware discriminator module, and cross-modal retrieval-augmented module were sequentially added, resulting in multiple model variants. Additionally, ablation experiments on key hyperparameters—including the number of retrieved samples and the physical guidance intensity—were conducted to validate the rationality of the parameter configurations.

From the module ablation results presented in Table 2, it can be observed that each innovative module contributes independent positive gains, and the combination of multiple modules exhibits significant synergistic optimization effects. With the introduction of the multi-level structural injection mechanism alone, structural fidelity metrics are substantially improved, demonstrating that the coarse-to-fine hierarchical constraint strategy effectively optimizes multi-scale geometric generation accuracy and fundamentally alleviates structural distortion problems. After the frequency-domain decoupled sampling strategy is superimposed, image realism metrics achieve a substantial leap forward, while structural metrics continue to exhibit modest improvements, indicating that the frequency-domain decoupling mechanism thoroughly eliminates feature coupling interference between low-frequency structures and high-frequency textures, refining textural details while preserving structural stability. The introduction of the physically aware discriminator further rectifies physical deviations in lighting and materials,

suppressing image artifact generation. Finally, with the addition of the cross-modal retrieval-augmented module, semantic alignment performance is significantly enhanced, while structural and textural generation accuracy are concurrently improved in a reciprocal manner, fully demonstrating that domain-specific prior information effectively compensates for the informational deficiencies of abstract inputs and achieves comprehensive performance optimization across all dimensions.

The hyperparameter ablation experiments shown in Table 3 indicate that an optimal balance range exists for both the number of retrieved samples and the physical guidance intensity. An excessively low number of retrieved samples leads to insufficient domain prior information, failing to effectively constrain the generation logic; an excessively high number introduces redundant features, resulting in style contamination and structural deviation. Insufficient physical guidance intensity fails to correct textural physical defects, whereas excessive guidance intensity disrupts the underlying spatial structural layout. The configuration adopted in this study—with K=5 and guidance intensity η=1.0—achieves an optimal balance among structural precision, semantic matching, and visual realism, demonstrating the scientific validity and rationality of the parameter configuration.

Table 2. Ablation experimental results of core modules

Model Variant	Edge Intersection over Union	Mean Intersection over Union	Fréchet Inception Distance	Inception Score	Contrastive language–Image Pre-Training Similarity
Baseline model	0.725	0.693	18.62	8.13	0.765
Variant 1 (+ multi-level structural injection)	0.789	0.756	15.87	8.42	0.781
Variant 2 (+ frequency-domain decoupled sampling)	0.805	0.773	12.65	8.96	0.803
Variant 3 (+ a physically aware discriminator)	0.812	0.781	11.28	9.25	0.827
Variant 4 (+ retrieval augmentation, a full model)	0.836	0.802	10.32	9.58	0.856

Table 3. Ablation experimental results of key hyperparameters

Retrieval Count K	Guidance Intensity η	Fréchet Inception Distance	Contrastive Language–Image Pre-Training Similarity	Mean Intersection over Union
3	0.8	12.15	0.832	0.785
5	1	10.32	0.856	0.802
7	1.2	11.06	0.849	0.796

3.4 Specialized experiments on multi-scale structural control precision

To precisely verify the adaptability of the hierarchical structural control mechanism across different structural scales, model structural accuracy was quantitatively evaluated at three levels—global layout, intermediate massing, and local details. Concurrently, robustness tests were conducted on three common designer input formats—sketches, layout diagrams, and wireframes—to comprehensively validate the model's adaptability across diverse scenarios.

As shown in Figure 6, the multi-scale structural experimental results indicate that traditional structural control models generally suffer from scale-adaptation imbalance, failing to simultaneously maintain both global layout integrity and local detail precision. The single-weight feature injection approach tends to lead either to the omission of detailed elements when global composition is stabilized, or to disorganized overall spatial layout when details are enriched. Through dynamic weight scheduling adapted to the generation requirements of different denoising stages, the proposed progressive hierarchical injection mechanism achieves optimal accuracy across all three scales—global spatial layout, intermediate architectural massing, and local landscape details. In the multi-input robustness tests, the model maintains high structural generation accuracy even when faced with hand-drawn sketch inputs characterized by sparse information and ambiguous features, demonstrating that the hierarchical structural encoding mechanism possesses excellent anti-interference capability and can accommodate diverse creative input forms from designers, with scenario adaptability significantly superior to that of existing methods.

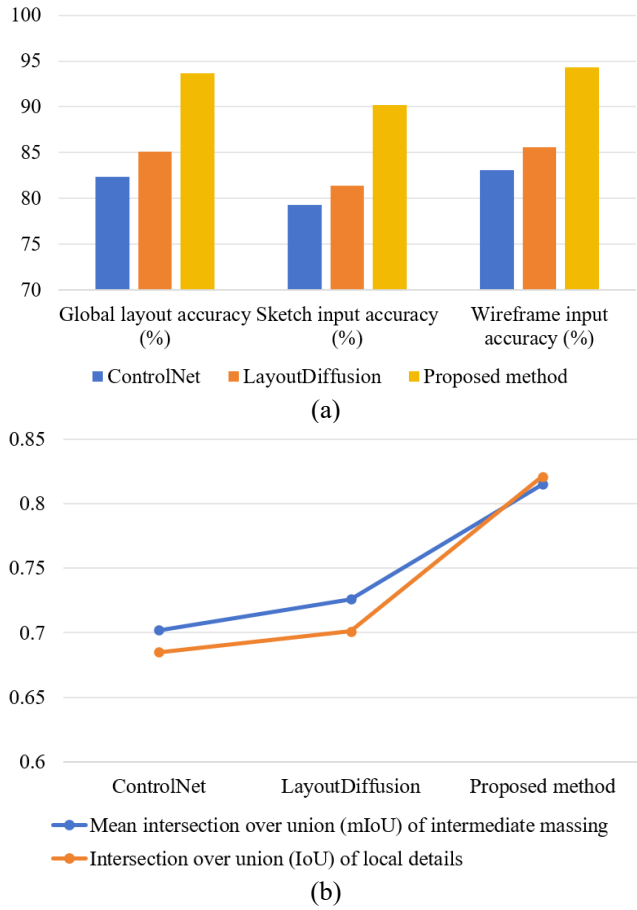

Figure 6. Experimental results of multi-scale structural control and multi-input robustness

3.5 Specialized evaluation experiments on physical realism

To precisely quantify the physical plausibility of the generated images, three dedicated physical evaluation metrics were constructed: illumination direction consistency error, shadow boundary softness deviation, and material bidirectional reflectance distribution function fitting error (Table 4), with lower values indicating higher adherence to physical laws. Combined with professional subjective blind evaluation scores, the optimization effectiveness of the physically aware discriminator was comprehensively assessed, with generation performance further analyzed across different material categories.

Table 4. Quantitative and subjective evaluation results for physical realism

Method	Illumination Error	Shadow Deviation	Bidirectional Reflectance Distribution Function Fitting Error	Subjective Realism Score
ControlNet	0.362	0.321	0.295	78.65
RealDiffusion	0.315	0.286	0.263	82.36
Proposed method	0.126	0.103	0.095	91.82

The physical realism experimental results demonstrate that general-purpose generation models and traditional realism enhancement models are incapable of precisely fitting the domain-specific physical rules of environment design scenarios, commonly suffering from disorganized illumination directions, distorted shadow morphology, and erroneous material reflectance properties. The proposed multi-dimensional physically aware discriminator enables refined constraints on the generation process from the three dimensions of illumination, shadow, and material, substantially reducing all physical deviation metrics. The model maintains unified global illumination direction, achieves precise matching between shadow scale and occluding objects, and effectively fits the optical reflectance characteristics of typical environmental materials—including stone, glass, and vegetation—without textural distortions or physical logic conflicts. The professional subjective evaluation scores are highly consistent with the quantitative results, demonstrating that the strategy combining frequency-domain decoupling optimization with physical gradient guidance significantly enhances the professional realism of environment design images and satisfies the physical realism requirements of industry visualization applications.

3.6 Robustness experiments on complex scenarios

To verify the stability of the model in high-difficulty design scenarios, robustness tests were conducted on three typical complex scenario types—large-scale site planning, complex irregular building facades, and high-density landscape layouts—with metric degradation rates, generation success rates, and structural failure rates statistically analyzed across models, thereby clarifying the scenario adaptation capability and applicability boundaries of the model.

The complex scenario experimental results presented in Figure 7 demonstrate that existing baseline models exhibit significant performance degradation when confronted with design scenarios characterized by complex topological structures and densely interleaved elements, frequently suffering from issues such as disorganized facade segmentation, overlapping landscape elements, and imbalanced site layouts, with poor generation stability. Relying on the multi-scale hierarchical constraint mechanism and domain retrieval-based prior constraints, the proposed framework effectively accommodates the generation requirements of complex scenarios, achieving a comprehensive metric degradation rate of only 3.21% in complex scenarios, a generation success rate of 96.85%, and a structural failure rate substantially lower than that of all comparative models. The model stably maintains overall spatial order in large-scale planning scenarios, while precisely preserving detailed structural features in irregular building and high-density landscape scenarios, demonstrating strong scenario robustness. The limited number of failure cases is concentrated exclusively in extremely complex irregular topological structures and exceptionally large-scale composite scenarios, where the geometric constraints are inherently highly challenging, thereby providing a clear research direction for subsequent iterative optimization of the model.

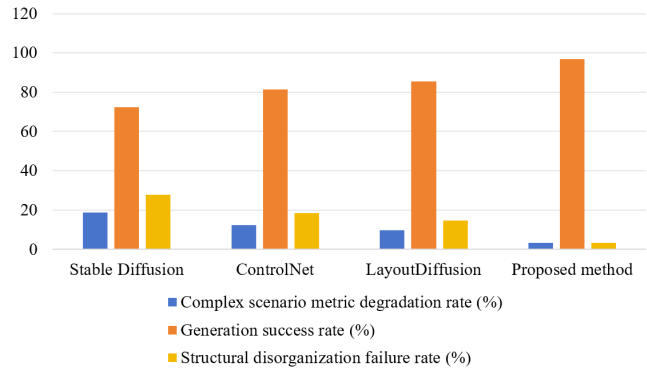

Figure 7. Robustness experimental results on complex scenarios

4. DISCUSSION

The StructDiff-Real framework constructed in this study effectively addresses the issues of structural inaccuracy, insufficient realism, and semantic drift commonly encountered in conventional environment design image generation. However, certain technical limitations remain under extremely complex design scenarios and engineering deployment conditions. When confronted with complex architectural and landscape structures featuring curved irregular surfaces, multiply nested topologies, and ultra-high-density components, the fine-grained geometric constraint capability of the model falls short of achieving fully precise replication, leaving room for improvement in the structural reconstruction accuracy of extreme details. Furthermore, the generation performance of the retrieval-augmented mechanism is highly dependent on the sample coverage breadth and annotation precision of the domain case knowledge base. In low-sample scenarios—such as niche regional design styles and customized innovative designs—the constraint efficacy of prior information exhibits notable

attenuation, thereby limiting the upper bound of the model's creative generation capacity. Additionally, although the latent-space frequency-domain decomposition operations and the gradient guidance mechanism of the physically aware discriminator enrich the generative constraint dimensions, they also increase the computational overhead of model inference to a certain extent, resulting in a modest inference latency compared to the base diffusion model, which renders it less suitable for the ultra-fast inference requirements of real-time interactive design rendering.

The core technical architecture of the proposed method possesses robust domain generalization capability and cross-task transfer potential, with each functional module not being confined to the single application scenario of outdoor environment design. The multi-scale structural hierarchical encoding, frequency-domain feature decoupling optimization, and cross-modal semantic alignment mechanisms introduced in this study can accommodate the general structural constraint requirements of spatial design tasks. Through merely updating the corresponding domain's structural annotation system and case knowledge base, seamless transfer can be achieved to related design domains—including interior spatial design and urban district planning—thereby enabling intelligent image synthesis across multiple scenarios. At the stylistic adaptation level, relying on the retrieval-augmented domain prior constraint system, the model adaptively accommodates the generation characteristics of diverse design styles—such as traditional Chinese, modern minimalist, and European classical—effectively balancing standardized design expression with differentiated stylistic creation, demonstrating excellent cross-style generation stability and providing a feasible technical pathway for the construction of general-purpose intelligent generation models for spatial design.

Targeting the industry pain points of low efficiency in traditional environment design visualization workflows and high costs of scheme iteration, this research has developed a practically deployable intelligent generation technology system with significant engineering application value and industrial empowerment potential. In practical design workflows, this method enables high-precision visual rendering of preliminary design drafts rapidly based on designers' lightweight sketch layouts and textual design descriptions, substantially shortening the turnaround time of conventional modeling and rendering processes. Concurrently, the model is capable of generating multiple structurally sound and stylistically differentiated scheme variants from a single core design logic, providing sufficient material support for designers to conduct scheme comparison and optimization, thereby effectively enhancing design iteration efficiency. This technology can be broadly applied to engineering scenarios such as architectural landscape scheme exploration, public space renovation design, and site planning visualization, assisting designers in completing preliminary creative realization and mid-stage scheme optimization, and promoting the digital transformation of the environment design industry toward intelligent and efficient paradigms.

5. CONCLUSION AND FUTURE WORK

Targeting the core technical deficiencies in environment design image generation—namely, imbalance in multi-scale structural constraints, lack of physical plausibility in materials and lighting, and susceptibility to semantic drift from abstract inputs—an integrated StructDiff-Real diffusion generation framework was constructed. Through three collaborative optimization modules, a complete generation pipeline was established, enabling simultaneous precise control over global layout, intermediate morphological structures, and local details via hierarchical structural encoding and progressive feature injection. Feature coupling interference between structural and textural representations was eliminated through latent-space frequency-domain decoupled sampling and a physically aware discriminator, while design semantics were stably constrained through the establishment of a domain knowledge base and a cross-modal alignment retrieval mechanism. Multiple sets of quantitative and qualitative comparative experiments, module ablation studies, and specialized performance tests demonstrated that the proposed method significantly outperformed all existing state-of-the-art models across all metrics—including structural fidelity, visual realism, and semantic matching—with the generated outcomes meeting the professional visualization standards of environment design. The research contributions not only complement the theoretical framework of controllable diffusion generation in professional design scenarios, but also provide a complete and feasible technical implementation pathway for rapid iteration of architectural and landscape design schemes.

Given the current performance boundaries and engineering deployment limitations of the proposed framework, three future research directions can be pursued. First, three-dimensional voxel-based geometric priors can be introduced into diffusion denoising optimization, moving beyond the limitations of two-dimensional planar representation to achieve controllable generation of complete three-dimensional structures with spatial metric properties. Second, the static image generation architecture can be extended to temporal diffusion models for the generation of continuous dynamic roaming videos of design scenes, thereby supporting immersive design scheme presentations. Third, through model distillation, parameter sparsification, and other lightweighting techniques, the computational overhead during the inference stage can be reduced, enabling the integration of the complete generation algorithm into various terminal design software platforms and facilitating real-time interactive visual creation for designers.

REFERENCES

[1] Boje, C., Guerriero, A., Kubicki, S., Rezgui, Y. (2020). Towards a semantic construction digital twin: Directions for future research. Automation in Construction, 114: 103179. https://doi.org/10.1016/j.autcon.2020.103179.

[2] Opoku, D.J., Perera, S., Osei-Kyei, R., Rashidi, M. (2021). Digital twin application in the construction industry: A literature review. Journal of Building Engineering, 40: 102726. https://doi.org/10.1016/j.jobe.2021.102726

[3] Omrany, H., Al-Obaidi, K.M., Husain, A., Ghaffarianhoseini, A. (2023). Digital twins in the construction industry: A comprehensive review of current implementations, enabling technologies, and future directions. Sustainability, 15(14): 10908. https://doi.org/10.3390/su151410908

[4] Delgado, J.M.D., Oyedele, L., Demian, P., Beach, T. (2020). A research agenda for augmented and virtual

reality in architecture, engineering and construction. Advanced Engineering Informatics, 45: 101122. https://doi.org/10.1016/j.aei.2020.101122

[5] Zhang, Y., Liu, H., Kang, S., Al-Hussein, M. (2020). Virtual reality applications for the built environment: Research trends and opportunities. Automation in Construction, 118: 103311. https://doi.org/10.1016/j.autcon.2020.103311.

[6] Safikhani, S., Keller, S., Schweiger, G., Pirker, J. (2022). Immersive virtual reality for extending the potential of building information modeling in architecture, engineering, and construction sector: Systematic review. International Journal of Digital Earth, 15(1): 503-526. https://doi.org/10.1080/17538947.2022.2038291

[7] Croitoru, F., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/tpami.2023.3261988

[8] Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

[9] Jiang, R., Zheng, G., Li, T., Yang, T., Wang, J., Li, X. (2024). A survey of multimodal controllable diffusion models. Journal of Computer Science and Technology, 39(3): 509-541. https://doi.org/10.1007/s11390-024-3814-0

[10] Yang, L., Huang, W. (2022). Representation and assessment of spatial design using a hierarchical graph neural network: Classification of shopping center types. Automation in Construction, 147: 104727. https://doi.org/10.1016/j.autcon.2022.104727

[11] Jo, H., Lee, J., Lee, Y., Choo, S. (2024). Generative artificial intelligence and building design: Early photorealistic render visualization of façades using local identity-trained models. Journal of Computational Design and Engineering, 11(2): 85-105. https://doi.org/10.1093/jcde/qwae017

[12] Lee, J., Yoo, Y., Cha, S.H. (2024). Generative early architectural visualizations: Incorporating architect's style-trained models. Journal of Computational Design and Engineering, 11(5): 40-59. https://doi.org/10.1093/jcde/qwae065

[13] Shi, M., Seo, J., Cha, S.H., Xiao, B., Chi, H. (2024). Generative AI-powered architectural exterior conceptual design based on the design intent. Journal of Computational Design and Engineering, 11(5): 125-142. https://doi.org/10.1093/jcde/qwae077

[14] Jiang, H., Luo, A., Fan, H., Han, S., Liu, S. (2023). Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics, 42(6): 1-14. https://doi.org/10.1145/3618373

[15] Huang, Y., Huang, J., Liu, J., Yan, M., Dong, Y., Lv, J., Chen, C., Chen, S., Lyu, J. (2024). WaveDM: Wavelet-based diffusion models for image restoration. IEEE Transactions on Multimedia, 26: 7058-7073. https://doi.org/10.1109/tmm.2024.3359769

[16] Li, Y., Shao, H., Liang, X., Chen, L., Li, R., Jiang, S., Wang, J., Zhang, Y. (2023). Zero-shot medical image translation via frequency-guided diffusion models. IEEE Transactions on Medical Imaging, 43(3): 980-993. https://doi.org/10.1109/tmi.2023.3325703

[17] Yu, Y., Smith, W.A.P. (2021). Outdoor inverse rendering from a single image using multiview self-supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3659-3675. https://doi.org/10.1109/tpami.2021.3058105

[18] Choi, J., Lee, S., Park, H., Jung, S., Kim, I., Cho, J. (2025). MAIR++: Improving multi-view attention inverse rendering with implicit lighting representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(6): 5076-5093. https://doi.org/10.1109/tpami.2025.3548679

[19] Bie, F., Yang, Y., Zhou, Z., et al. (2024). RenAIssance: A survey into AI text-to-image generation in the era of large model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 2212-2231. https://doi.org/10.1109/tpami.2024.3522305

[20] Cao, Y., Aziz, A.A., Arshard, W.N.R.M. (2024). Stable diffusion in architectural design: Closing doors or opening new horizons? International Journal of Architectural Computing, 23(2): 339-357. https://doi.org/10.1177/14780771241270257