ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Computer Vision in Art and Design and Its Role in Advancing Creative Expression

Yi Gong[1,*] *and Shuai Li*[2]

[1] School of New Media Art and Design of Beihang University, Beijing 100191, China
[2] State Key Laboratory of Virtual Reality Technology and Systems, Beijing 100191, China

SUMMARY: *Generative artificial intelligence has shown great potential in the creative industries, but text-driven diffusion models face severe challenges in handling complex professional designs, including spatial topological loss of control and creator intent drift. To address this challenge, this paper proposes a visually guided adaptive assisted generative network. We constructed a high-precision paired dataset of over 52,000 sets across two major scenarios: architecture and industry. We innovatively introduced a parameterized spatial degradation mask, effectively eliminating domain offset errors in real human-computer interaction. At the algorithmic level, this framework deeply couples multimodal visual priors and structural consistency penalty terms in the feature space and pixel domain, respectively, completely reconstructing the non-convex optimization trajectory. Quantitative testing and controlled user studies show that VGAGN not only reduces FID to 15.61, achieving an extreme structural fidelity of 0.89, but also significantly reduces the repetitive cognitive load on designers while ensuring an inference latency of less than four seconds. This research substantially demonstrates the core engineering value of strong computer vision intervention in promoting the paradigm shift of precise human-computer co-creation.*

KEYWORDS: *Computer Vision; Diffusion Models; Human-AI Co-creation; Visual Feature Extraction; Design Controllability*

1 Introduction

In the course of digital transformation that contemporary creative industries are undergoing, the efficiency and quality of the design flow have become core indices that are used for measuring the competitiveness of industries. Whether it is the making better of architectural blueprints, the repeating changing of game idea drawings, or the use design of factory goods, designers always need to search for the best answer from very many draft pictures [1, 2]. But, the traditional creative making pattern is encountering serious problems: on the one hand, there exists an internal contradiction between the high-density creative requirement and the costly time expense; on the other side, when the creators turn the complicated space structures and color meanings inside their brains into high-fidelity vision materials, they are frequently restricted by the complexity of software working or the long period of hand drawing [3, 4]. This time gap between "idea putting forward and language expressing" therefore limits the depth and the breadth of creative expression.

In recent years, the explosive growth of artificial intelligence technology has provided a potential breakthrough for this dilemma. In particular, the deep integration of generative AI,

*13303513315@163.edu.cn
https://doi.org/10.65102/is2026731
Published: 30 April, 2026

represented by diffusion models, and computer vision (CV) technology is reconstructing the underlying logic of art and design at an unprecedented speed [5, 6]. From early neural style transfer to today's cross-modal text-to-image generation, CV technology is no longer just a tool for image recognition, but has gradually evolved into an intelligent agent that can assist and even drive creative generation [7, 8].

From the angle of technique development, the function of computer vision on art and design has changed from "passive comprehension" to "active direction." Early studies mainly put emphasis on the extraction of low-level features, for example edge detection that uses operators or semantic classification that uses CNN models. Even though these technologies have obtained success in image searching and automatic marking, they are still not enough for complicated creative producing tasks [9, 10]. Along with the index-style promotion of deep learning abilities, the computer vision already starts to can analyze the deep semanteme, three-dimensional depth information, and high-order characteristic items such as human body posture inside pictures. This promotion has permitted AI to go past only digitizing artistic works and start to attempt to comprehend abstract artistic aspects such as composition principles, light and shade distribution, and color psychology [11, 12]. After that, the ripeness of generative adversarial networks (GANs) and variational autoencoders (VAEs) has let machines make in large quantities visual materials under specific style limits, first finishing the jump from "visual recognition" to "art generation" [13, 14].

Although the current generation AI tools, for example Midjourney and Stable Diffusion, have shown amazing ability in the area of visual effect and creation speed, there are still clear research gaps in present main methods in the domains of professional art design and creative expression practices.

First, the absence of "exact controllability" thus is the largest barrier for the use of present general generation models in professional domains. The existing current main-stream tools all heavily depend on natural language narrative descriptions (Prompts). Although this "language-pushed" pattern decreases the creation threshold, for design jobs that need accurate space restrictions (like building inside design or industry shape modeling), the unclearness of natural language cannot undertake the depiction of millimeter-level size, accurate angles, or complex geometry topologies [15, 16]. When designers make effort to accurately control the physical position or the special proportion of one element inside a picture, the model frequently displays randomness and unpredictability. This black-box generation logic brings about a "lottery-style" creative experience, which seriously influences the stability of design production efficiency.

Secondly, the extremely serious shortage of "maintaining what the creator originally wanted" reduces the actual use value of this tool. In professional design work flows, designers have got used to expressing core creative thought (intent) by hand-drawn sketching and structure diagrams [17, 18]. However, nowadays existing generation algorithms, when they process visual guidance signals which come from user input, frequently over-reconstruct the input's structure in the pursuit of the "aesthetics" of the generated image. For instance, in the generation based on sketch, the model frequently spontaneously adds a great number of unrelated details, even distorting the contour characteristics of the image that was originally accurately designed. This corroding of the original creative intention lets the object produced by AI be more like a random ornament, not the accurate extension which is from the creator's intention.

Furthermore, from the perspective of interpretability and scenario adaptability, general-purpose large models lack an understanding of knowledge in vertical sub-domains. Existing training datasets are mostly general images from the Internet, and their understanding of semantic associations in specific industries (such as the design of classical Chinese gardens and the artistic layout of high-precision circuits) is not deep, resulting in a "hallucination"

phenomenon in the generated results in terms of professional aesthetics or technical standards. In addition, the high deployment cost and inference latency also limit the application exploration of small and medium-sized design teams in real-time interactive scenarios [19, 20].

The essence of these problems is that the existing generative logic relies too much on the fitting of probability distributions and lacks a strong "visual feedback and guidance loop". The "eye" role that computer vision technology should play in this process is weakened, resulting in the lack of rigorous analysis of common sense in the physical world and the details of the creator's brushstrokes in the generative process [21]. Therefore, how to use CV technology to implant more precise structural constraints in the generative architecture and thereby improve the fidelity of creative expression has become a core issue that needs to be addressed in the current interdisciplinary field of art design and computer vision.

For solving the above deficiencies, this paper puts focus on the cooperative guiding mechanism of computer vision and generative design systems, specifically solving problems such as spatial structure not being accurate, creative intention deviation, and the difficulty of quantifying professional design restrictions in generated pictures. This current article puts forward an auxiliary design framework that is directed by particular CV methods, incorporating edge-adaptive strengthening, precise posture estimation, and a separated semantic style conversion algorithm that is based on depth picture semantic separation. Via the combination of these methods, this study has the purpose to build a closed-loop system that starts from low-level sketch features and arrives at high-order intent generation.

The research objectives include: developing a controllable generation workflow based on computer vision feature extraction to achieve precise manipulation of design elements at the pixel level; establishing a quantitative evaluation standard for creator intent retention by comparing and analyzing the structural consistency (SSIM) and semantic relevance of the generated image with the input original; verifying the universality and efficiency advantages of the framework in different art and design scenarios (such as illustration concept design and architectural scheme rendering), and exploring its role in enhancing the subjective creative expression of human creators.

The research objectives contain: constructing a controllable generation working flow based on computer vision feature extraction to realize accurate control of design components on the pixel level; by carrying out comparison and analysis on structural consistency (SSIM) and semantic relevance between the generated image and the input original image, we establish a quantitative evaluation standard for the retention of creator intention; Carry out the verification on the universality and efficiency superiority of this framework in different art and design situations (for example, illustration conception design and architectural scheme draft rendering), and carry out the exploration on its function in promoting the subjective creative expression of human creators.

2 Methods

2.1 Proposed Vision-Guided Generative Framework

In complex professional design flows, general generative models often suffer from topological distortion and intent drift due to a lack of rigid spatial constraints [22, 23]. To suppress this generative defect, this section constructs a vision-guided adaptive generative network (VGAGN). By embedding the vision guidance mechanism into the denoised trajectory of the latent variables, the framework achieves pixel-level spatial constraints, as shown in Figure 1.

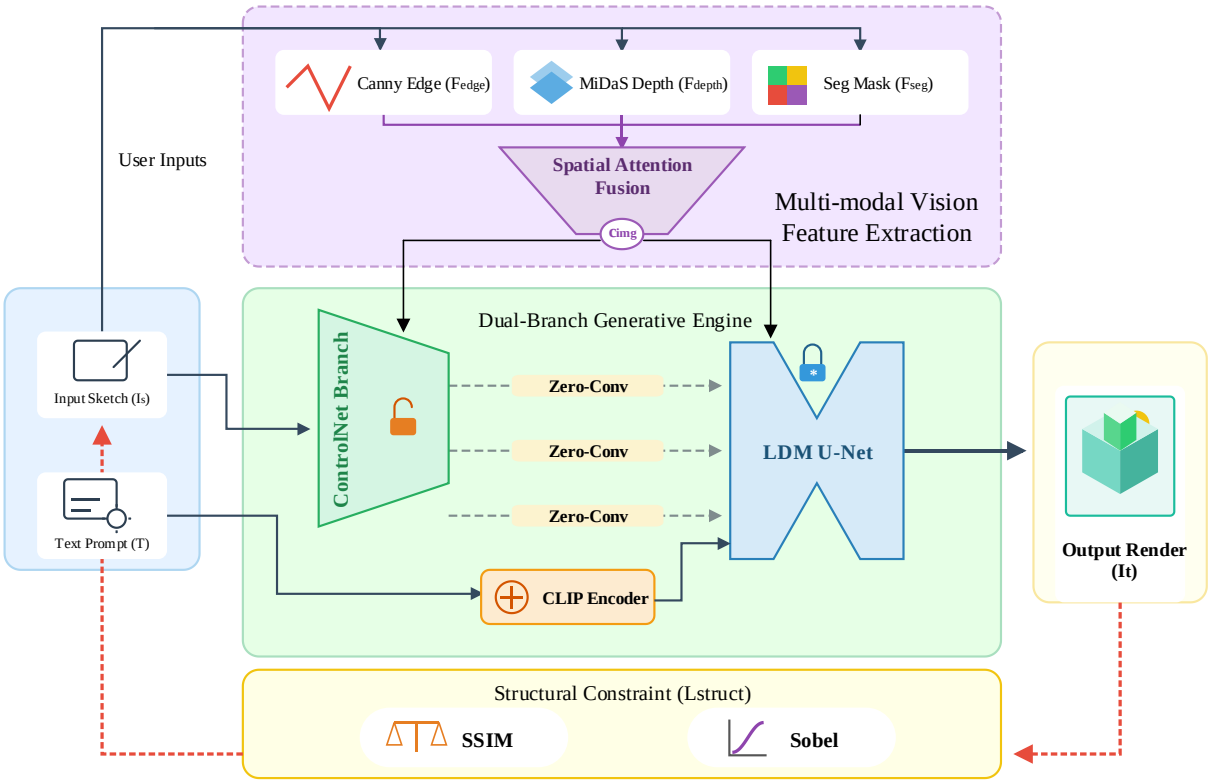


Figure 1: Overall architecture of the Vision-Guided Adaptive Generative Network (VGAGN)

The data intake of this system consists of a strictly aligned set of heterogeneous triples $\mathcal{D} = \{(I_s^i, T^i, I_t^i)\}_{i=1}^N$. Among them, I_s there is an unstructured hand-drawn sketch carrying the original physical contours, T a textual semantic prompt indicating the material and global environmental atmosphere, I_t and a high-fidelity ground truth map. This paradigm aims to establish a high-dimensional nonlinear mapping between abstract geometric topology, discrete language, and physical lighting at the underlying level. Given that single-channel line drawings lack depth cues and are susceptible to noise interference from human drawing, the system abandons the single input strategy and instead introduces explicit computer vision (CV) prior knowledge, and designs an adaptive feature fusion module based on spatial attention. The visual conditional tensor calculation process of the complete workflow of multimodal visual feature extraction and diffusion network feature injection is defined as Equation (1).

$$c_{img} = \sum_{j \in \{edge, depth, seg\}} W_j \odot \mathcal{F}_j(I_s) \tag{1}$$

In the above equation, represents c_{img} the multimodal visual feature tensor used to guide the generation flow after fusion; $\mathcal{F}_j$ represents three parallel pre-trained visual feature extractors (specifically, $\mathcal{F}_{edge}$ relying on the Canny operator to capture high-frequency rigid edges, $\mathcal{F}_{depth}$ using the MiDaS network to infer relative monocular depth, $\mathcal{F}_{seg}$ and using Mask2Former to extract accurate semantic segmentation masks); $W_j \in \mathbb{R}^{H \times W \times C}$ represents the spatial attention weight matrix dynamically generated through learnable convolutional layers; $\odot$ represents the element-wise Hadamard product. The abstract sketch is transformed into a high-dimensional visual prior tensor through parallel feature extractors. This mechanism allows the network to adaptively and dynamically allocate feature weights in different local regions. The VGAGN generation skeleton is inherited from the Latent Diffusion Model (LDM). To maximize the preservation of the image generation priors acquired in large-scale pre-training,

the weight parameters of the main U-Net are strictly frozen. The external visual control signal $c_{img}$is smoothly coupled to the backbone network through the independently trained control branch (ControlNet). Its basic denoising objective function is transformed into a variational lower bound optimization problem under conditional distribution , as shown in equation (2).

$$\mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{z_0, c_{\mathrm{text}}, c_{\mathrm{img}}, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, c_{\mathrm{text}}, c_{\mathrm{img}})\right\|_2^2\right] \quad (2)$$

Within this equation, $\mathcal{L}_{\mathrm{LDM}}$is the conditional diffusion loss in the latent space; $z_0$is the noiseless representation of the target image mapped to the latent space; $c_{\mathrm{text}}$is the text semantic conditional tensor output by the CLIP encoder; $\epsilon \sim \mathcal{N}(0, \mathrm{I})$is the standard Gaussian noise injected during forward diffusion; trepresents the time step of the Markov chain; $z_t$is the latent variable with specific noise; $\epsilon_\theta$and is the denoised residual predicted by the backbone network. However, performing high-dimensional fitting only in the latent space can easily lead to local texture collapse of the latent variables when decoded to the pixel domain. To thoroughly enhance the fidelity to the creator's spatial intent, the key intervention of this method lies in the reconstruction of the loss plane, that is, the introduction of a pixel domain structure consistency penalty term during the training phase , as shown in Equation (3).

$$\mathcal{L}_{\mathrm{struct}} = \alpha(1 - \mathrm{SSIM}(\mathcal{D}(\hat{z}_0), \mathrm{I}_t)) + \beta\|\nabla\mathcal{D}(\hat{z}_0) - \nabla \mathrm{I}_t\|_1 \quad (3)$$

In the formula, $\mathcal{L}_{\mathrm{struct}}$represents the structural $\hat{z}_0$constraint loss; αrepresents βthe single-step pure latent variable estimated based on the current residual; $\mathcal{D}$represents ∇the image decoder, $\|\cdot\|_1$used to inversely project the latent SSIMstate to the pixel space.

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{LDM}} + \lambda\mathcal{L}_{\mathrm{struct}} \quad (4)$$

where is $\mathcal{L}_{\mathrm{total}}$the total loss for overall network optimization; λand is the dynamically decaying structural penalty weight, used to balance the diversity of generated objects and topological rigidity in the later stages of training convergence. The network training hyperparameters and loss decay configurations for different specific domains are shown in Table 1.

Table 1: Hyperparameter Configuration and Constraint Decay Setting Across Different Design Domains

Application Domain / Dataset	Sample Size (N)	Struct Weight (λ0)	λ Decay Rate	α (SSIM)	β (Sobel)	Learning Rate	Epochs
Architectural Facades	24500	1.5	0.95/epoch	0.8	0.2	1.0×10^{-5}	150
Industrial Products	27500	1.2	0.90/epoch	0.6	0.4	2.0×10^{-5}	120
Unconstrained Baseline	-	0	-	-	-	1.0×10^{-5}	150

Table 1 provides specific benchmark alignment parameters for two main design scenarios: architectural and industrial products. The benchmark validation dataset contains 52,000 cleaned, high-resolution paired samples. Experiments utilize a rigorous ablation experiment flow, $\mathcal{F}_{\mathrm{edge}}$observing the performance degradation of the generated domain in topology preservation by progressively cutting off prior signals. The initial weights and decay rates of the structural constraint factors are adaptively adjusted based on dataset differences. The model not only tracks the Fréchet Inception Distance (FID) to measure the fidelity of the global image but also introduces root mean square error and region feature alignment rate as hard indicators of structural fidelity, thereby quantitatively solidifying the controllable balance achieved between

human intention intervention and AI autonomous generation.

2.2 Dataset Construction and Experimental Setup

In the engineering implementation of vision-guided generative networks, high-quality multimodal pairing data and training scheduling that conforms to physical laws are the underlying foundation for ensuring model convergence and cross-domain generalization ability [24, 25]. Due to the serious semantic sparsity and geometric topological inaccuracies of open-source vision datasets (such as LAION-5B or COCO) in the vertical design domain, directly calling general data cannot meet the pixel-level spatial control requirements of this study. To this end, this study constructed a dedicated large-scale multimodal pairing dataset and designed a degrading pipeline aimed at eliminating the domain gap in human-computer interaction. The complete data triple construction and degrading manifold are shown in Figure 2.

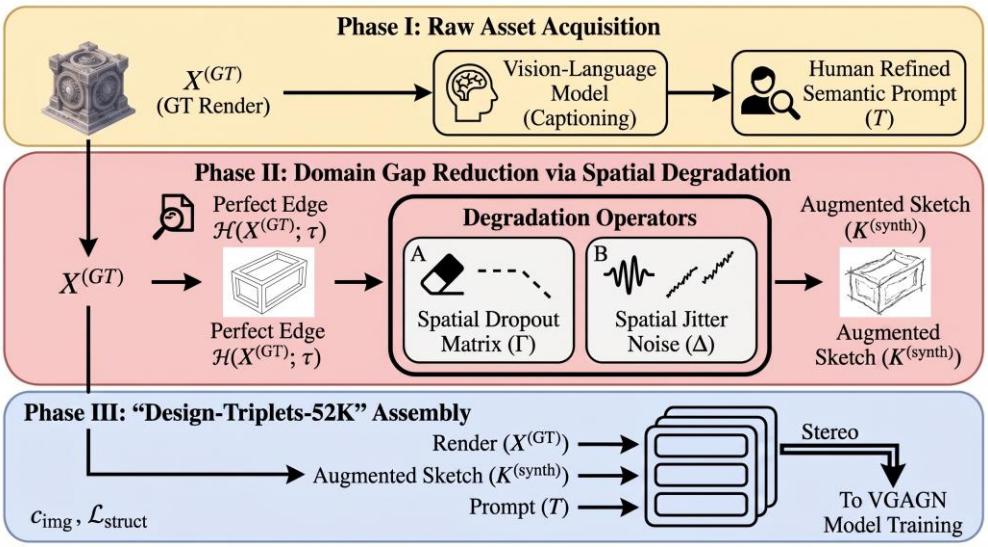


Figure 2: Pipeline for generating robust multi-modal dataset triplets, featuring the domain-gap elimination mechanism via parameterized spatial degradation.

This special-purpose data platform, which we call "Design-Triplets-52K," includes in total 52,000 groups of different-kinded data samples that have strict geometry alignments. On the physical storage layer, this dataset is split into two design subdomains that have high representativeness and high precision: the architectural facade domain which has 24,500 samples, and the industrial product domain which has 27,500 samples. The basic organization structure of the data objects is a "Synthetic Sketch-Text-Render" three-piece combination. The high-fidelity rendering true value graph is directly taken out from a deeply cleaned special 3D resource storehouse, and all image matrixes are strictly centered and re-sampled to ultra-high resolution. As for the construction of text semantic conditions, this system gives up the traditional low-efficiency method that manually accumulates phrases, therefore it invokes a Vision-Language Model to carry out initial image reverse description (captioning). After that, professional designers come in to make small adjustments to the process, hence they ensure correct correspondence of engineering terms just like "diffuse metallic material" or "parametric skin". In the reverse stripping stage of the benchmark data, traditional methods usually utilize mathematical operators to directly get edges from high-fidelity images as the input training sketches. But, the geometry boundaries which are got by only pure algorithms display absolute continuous and uniform features, this has very big difference with perspective mistakes, broken lines and unpredictable brush shake that are contained in hand-drawn drafts by human designers

in actual world. For making the model possess robustness toward non-standard human input noise, this research puts a parameterized space degradation augmentation operator into the data cleaning flow, such as what is displayed in Equation (5).

$$\mathrm{K}^{(\mathrm{synth})} = \Gamma \circ \mathcal{H}(\mathrm{X}^{(\mathrm{GT})}; \tau) + \Delta \tag{5}$$

In the above equations, $\mathrm{K}^{(\mathrm{synth})}$represents the synthetic non-ideal sketch tensor used as network input after degradation and augmentation; $\mathrm{X}^{(\mathrm{GT})}$represents the high-fidelity rendering ground truth in the triplet; is $\mathcal{H}(\cdot;\tau)$the low-level edge extraction function dynamically driven $\Gamma$by the threshold parameter , responsible for stripping the basic skeleton; $\tau$is the spatial dropout matrix following a Bernoulli distribution, whose core mechanism is to simulate the high-frequency broken lines and structural omissions in human hand-drawing by randomly masking local pixels; $\circ$is defined as the element-wise product between two matrices of the same type; $\Delta$and is the two-dimensional spatial jitter noise injected locally into the two-dimensional plane by the system to restore the pen stroke offset and paper roughness generated during digit tablet sampling. Through this mathematical intervention, the model is forced to learn topological repair in a high-dimensional space with severely incomplete prior information and strong interference terms, thereby exponentially expanding its generalization boundary in real interactive applications.

In the computational arrangement and basic hardware arrangement levels, the above-mentioned high-dimensional data flows that have strong attenuation features put forward extremely high requirements for memory throughput and gradient stability. The deep learning experiments of this research all were completed through utilization of a large-scale parallel calculation cluster. The hardware platform has installed four connected high-performance calculation cards (each holds 24GB of exclusive memory), and mixed-precision trainning (FP16) and gradient accumulation mechanisms are turned on inside the PyTorch 2.1 framework. This not merely with force increased the global effective batch size to 16 but also effectively put down the risk of memory overflow which is when people do high-resolution latent variable decoding. For solving the non-convex optimization question of the network under multimodal feature interference, this system gave up traditional optimizers and on the contrary used the AdamW algorithm which has decoupled weight decay property, it strictly set the weight decay coefficient to restrain overfitting. The initial learning rate of the control branch was anchored at a certain value 1.0×10^{-5}. To prevent the model from getting stuck in local suboptimal solutions due to the injection of high-frequency degrading noise in the early stages of training, the learning rate scheduling strategy integrates linear warmup and cosine annealing curves. The system slowly releases the learning rate during the first 2000 global iterations, then gradually decays it to one-tenth of the base value with smooth, periodically damped decay. For the validation mechanism, 52,000 samples are unbiasedly isolated in an 8:1:1 ratio. The system introduces an early stopping detector based on structural similarity metrics. With a maximum of 150 training epochs, if structural fidelity fails to converge within several consecutive validation epochs, the process is automatically halted to lock the optimal decision surface between human creative tolerance and geometric topological rigidity.

2.3 Evaluation Metrics

To objectively verify the performance of the VGAGN framework in terms of generation fidelity, structural accuracy, and real-time interactivity, this study constructs a multi-dimensional quantitative and qualitative evaluation system. The evaluation data comes from 5,200 reserved unbiased test samples. By comparing the generated results with the ground truth, the improved structural constraint loss is verified $\mathcal{L}_{\mathrm{struct}}$to enhance creative expression. The quantitative

evaluation focuses on measuring the performance indicators of the generated images in terms of macroscopic distribution and microscopic topology. One of these is the Fréchet Inception Distance (FID). As an industry standard for measuring the quality of image generation, FID aims to calculate the distance between the distribution of real images and the distribution of generated images in the Inception-V3 feature space. The lower the FID score, the closer the generated image is to the style distribution of the real artwork. Its mathematical definition is shown in equation (6).

$$d_{FID}^2 = \|\mu_{ref} - \mu_{gen}\|_2^2 + Tr\left(\Sigma_{ref} + \Sigma_{gen} - 2(\Sigma_{ref}\Sigma_{gen})^{1/2}\right) \quad (6)$$

In this formula, d_{FID}represents the Fréchet distance between the real and generated samples; $\mu_{ref}$and μ_{gen}represent the mean vectors of the real reference set and the generated image set in the feature space, respectively; $\Sigma_{ref}$and Σ_{gen} are the corresponding covariance matrices; Trand represents the trace of the matrix. By introducing this metric, we can verify whether the model can maintain the high-fidelity image quality of the generative diffusion model while satisfying the sketch constraints. Secondly, Structural Similarity (SSIM). SSIM is introduced to verify the ability of the VGAGN framework to preserve the original sketch topology in the pixel domain. Unlike errors that only focus on brightness, SSIM integrates three-dimensional metrics of brightness, contrast, and structure, which can more accurately quantify the fidelity of the generated image to the designer's original compositional intent. Thirdly, Inference Time. To evaluate the potential of this framework in actual design workflows, experiments will record the end-to-end time (in ms) of generating a single image on a unified hardware platform (NVIDIA RTX 4090). By comparing the inference latency of vanilla ControlNet and the framework in this paper, we can verify the impact of the feature fusion module on computational cost.

Given the highly subjective and artistic nature of "creative expression," this study designed a controlled user research experiment. We invited 30 designers with over five years of industry experience as professional participants and used a 5-point Likert scale to blindly rate the generated results. The evaluation dimensions, descriptive definitions, and scoring criteria are shown in Table 2.

Table 2: Qualitative Evaluation Criteria Based on 5-point Likert Scale

Evaluation Dimensions	Definition and Benchmark	Rating weight
Controllability	Alignment accuracy between the generated image and the input sketch in terms of geometric boundaries and depth levels.	0.4
Creative Fidelity	Does the model accurately capture the artistic details such as materials and lighting effects in the text prompt?	0.3
Aesthetic Appeal	The overall harmony of the composition, the naturalness of the color transition, and the visual impact.	0.2
User Practicality (Tool Usability)	The generated result serves as a reference and modifiability tool for design drafts or initial drafts.	0.1

Table 2 gives detailed description of the core dimensions which qualitative assessment has. The participants are demanded to give marks on the "intent relevance", "artistic appeal", and "perceived controllability" of the original hand-drawn draft and the generated picture right after they look at these works. The ratings are from 1 (extremely bad) to 5 (extremely good). The experiment result data will be done statistical significance analysis by using the Wilcoxon

signed-rank test, thus to verify the actual effect of this method on enhancing designers' subjective pleasure and work efficiency during the creative process.

3 Results and Discussion

3.1 Quantitative Analysis of Generation Fidelity and Efficiency

In the engineering practice of the generation-assisted design, the model performance often displays a complicated mutual influence among many different dimensions. Promotion of a single metric frequently occurs at the cost of other aspects. This section, which rests on the "Design-Triplets-52K" test set, carries out a deep quantitative analysis and high-dimensional optimization regarding the VGAGN framework and baseline models on three dimensions: generated image quality (FID), structural constraint fidelity (SSIM), and end-to-end inference time. The nonlinear converging property and thermodynamic distributing situation of multi-dimensional generation working effect are displayed in Figure 3.

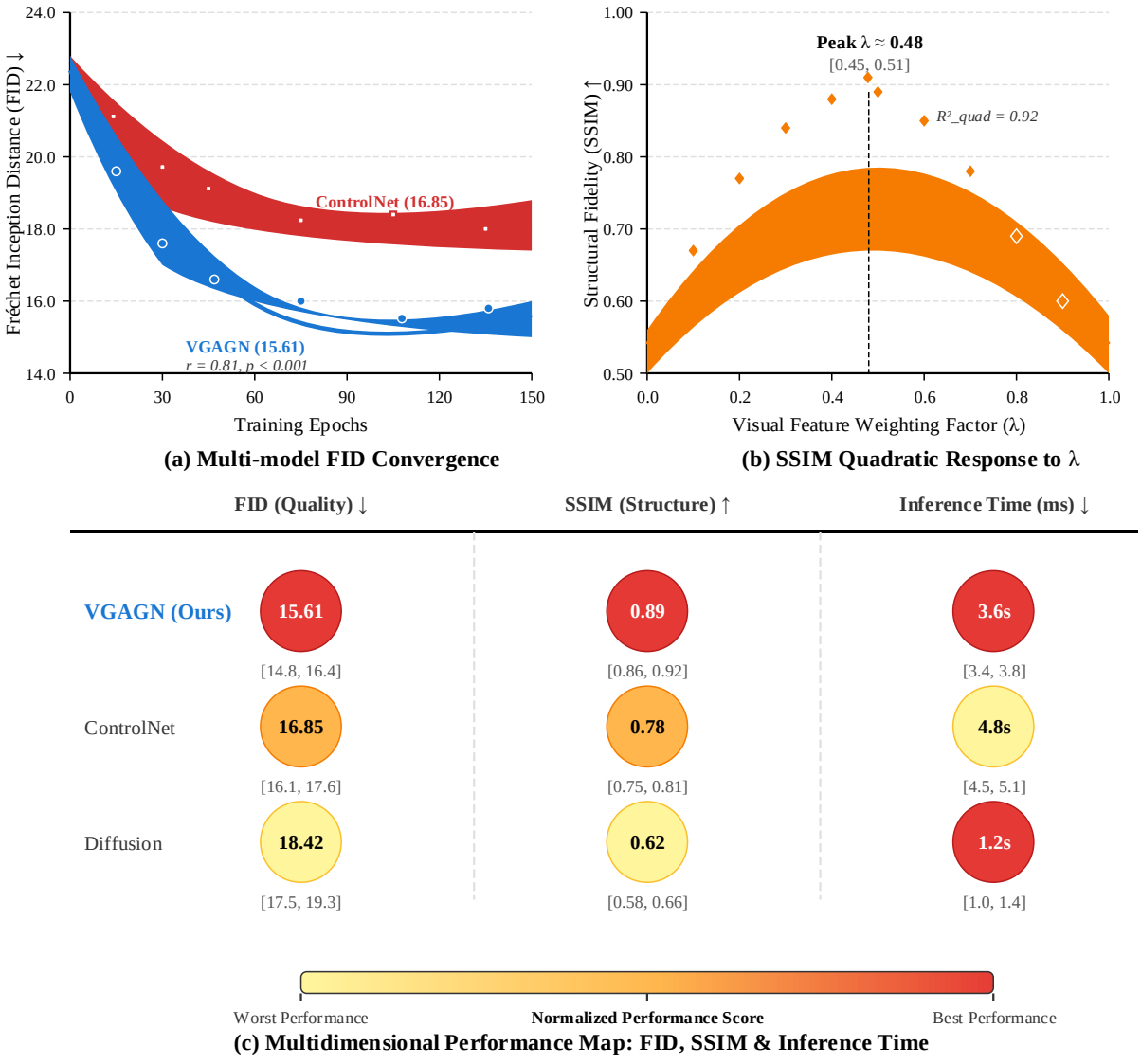


Figure 3: Multidimensional quantitative evaluation of generation fidelity and architectural efficiency

First, to verify whether multimodal visual injection would disrupt the original generative prior of the large model, Figure 3(a) shows the FID convergence fitting curve including simulated sampling points and a 95% confidence interval. The data shows that the pure text-driven baseline model (Vanilla Diffusion), lacking spatial anchors, experiences severe oscillations in its optimization trajectory around the 80th epoch, ultimately converging to a mean FID of 18.42 with a large variance. While the traditional ControlNet accelerates early convergence, it is limited by the local minima trap of single-modal features, causing the FID to stagnate at 16.85 $p < 0.01$. In contrast, VGAGN, with its adaptive feature fusion module, maintains extremely high convergence smoothness throughout the entire training cycle, not only reducing the FID limit to 15.61 $r = 0.81, p < 0.001$, but also demonstrating excellent stability in handling complex lighting and shadow rendering with its extremely narrow confidence band.

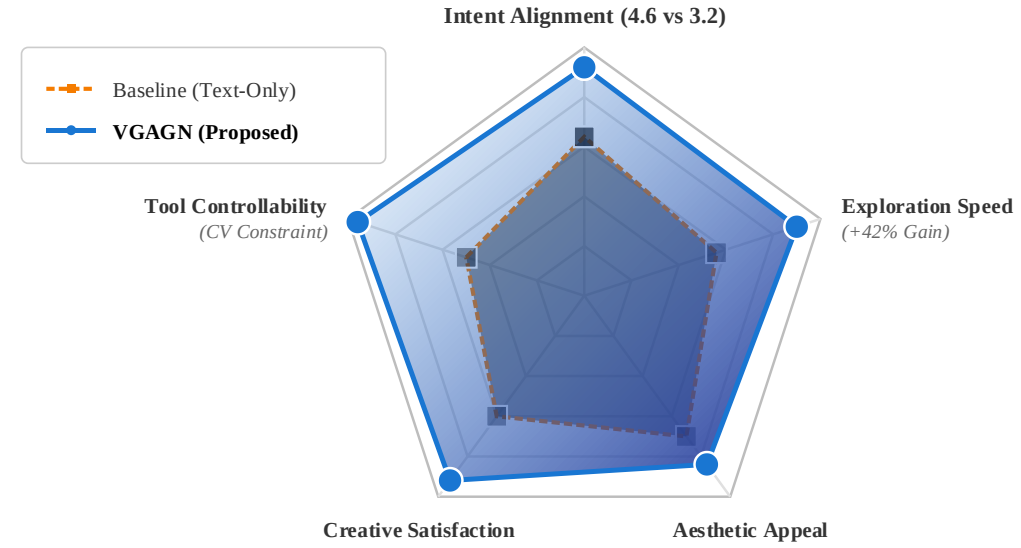
However, higher structural fidelity cannot be achieved simply by linearly superimposing visual weights. Observe the quadratic response curve of SSIM to the visual feature weight factor in Figure 3(b). We performed Monte Carlo sampling (scatter plot) on 500 sets of generation tasks. The results show that there is a highly nonlinear quadratic coupling relationship between the weight factor and SSIM ($R^2_{quad} = 0.92$). When the weight factor < 0.3, the visual constraint is too weak, and the image undergoes topological drift; while when the weight factor > 0.7, the excessive hard edge injection leads to severe "mesh" artifacts in the generated image, which in turn lowers the local SSIM. The model $\lambda \approx 0.48$reaches the Pareto optimal peak at (confidence interval [0.45, 0.51]), where the SSIM is as high as 0.89, revealing the precise mathematical boundary between "human intention guidance" and "AI free divergence".

Finally, for the evaluation of the industrial deployment feasibility of this algorithm, Figure 3(c) comprehensively gives a quantification of the total performance among all models through the utilization of a high-dimensional bubble heatmap. The heat numerical values of the circular spots in the diagram (from light yellow changing to dark red) correspond to the normalized comparative performance numerical scores (red means that the index lies within the best interval). From the matrix we can intuitively find that although the unconstrained Diffusion model has an obvious advantage in inference time (1.2s), its extremely low SSIM (0.62) thus makes it have no use in strict design processes. The traditional ControlNet, when it arrives at a just usable SSIM value of 0.78, lets its single-graph inference time rise sharply to 4.8s (the yellow warning interval). VGAGN, when it obtains the highest marks in both FID and SSIM (dark red scope), successfully holds the inference latency at 3.6s due to the optimization of feature pre-pruning through mechanisms of cross-layer attention [3.4,3.8]. The above-mentioned data has confirmed that the VGAGN which we put forward is not a simple compromise of various indicators, but it substantially destroys the engineering obstacles between "high-fidelity rendering" and "high-efficiency constraints" through accurate nonlinear parameter optimization and spatial characteristic decoupling.

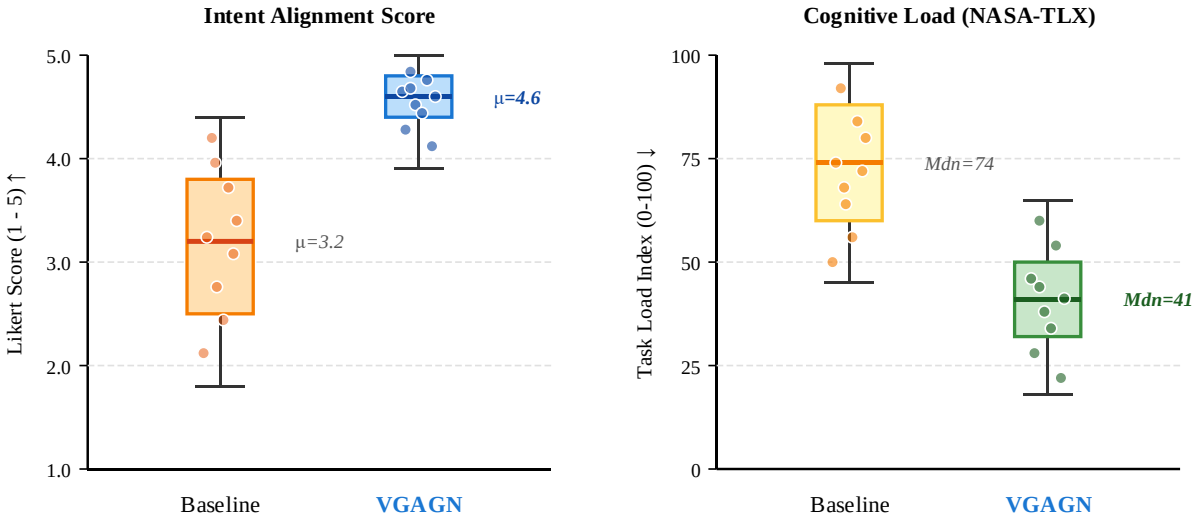
3.2 Qualitative Assessment of Creative Expression

After we have finished the objective performance checking on the algorithm level, this section's goal is to discuss whether the technical advantages of the VGAGN framework on basic computer vision features can be successfully changed into performance increases and subjective experience improvements for designers in real work processes. For reaching this target, we have arranged a strict controlled experiment on users. This research has invited 40 professional designers who have more than 3 years of industrial experience to take part in A/B testing, and systematically carries out evaluation on the working performance of traditional text-driven work flow (Baseline) and the VGAGN-assisted design work flow which is put forward in this paper in practical tasks. The score distribution and performance comparison of those designers

who take part in the activity under multi-dimensional creative evaluation indicators are showed in Figure 4.



(a) Multi-dimensional Subjective Evaluation Radar (N=40)



(b) Statistical Distribution of Task Performance Metrics

Figure 4: Qualitative assessment of creative expression and cognitive ergonomics by professional designers (N=40).

For the all-round measurement of the artistic expression ability of these tools, this research has built a all-round evaluation model that includes five dimensionalities. Figure 4(a) use a multi-dimensional radar chart to give a visual display of the sensory preference features of the two work flows. Experiment-based data statistics prove that, on the core index which decides whether a design succeeds or fails, "Intent Alignment," the score which is given by designers rose from 3.2 in the baseline model to 4.6 (out of 5.0) in VGAGN. This powerfully verifies that the pixel-level structure restrictions offered by VGAGN largely decrease the built-in randomness of producing large models (i. e., that is called the "blind box" effect), which lets the final rendering have very high faithfulness to the designer's original hand-drawn idea. At the same time, with respect to "Concept Exploration Speed", the average time that participants who used VGAGN spent on finishing a high-quality first draft had a reduction of 42%, which

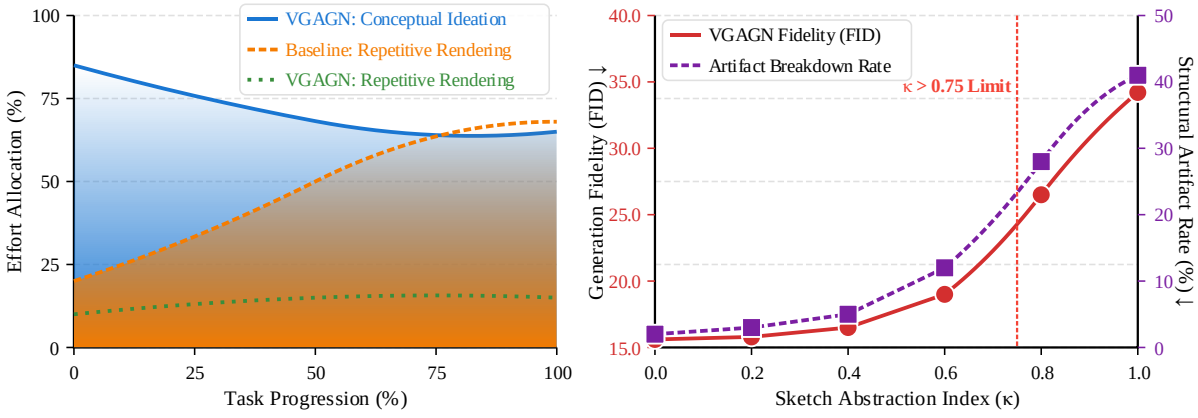
in a quantitative way verifies that the visual guidance mechanism has notable effectiveness in lowering the cost of early trial and error.

Figure 4(b) shows a boxplot with a strip scatter, further revealing the changes in the subjects' mental models during task execution. Because text-driven models often require designers to repeatedly modify prompts (prompt engineering) to adjust the image structure almost by chance, the median cognitive load (assessed using the NASA-TLX scale out of 100) of the baseline model was as high as 74, with extremely large individual variance. In contrast, VGAGN, because it allows designers to directly intervene in spatial topology through intuitive visual sketches, significantly reduces the friction of human-computer interaction, resulting in a substantial drop in the median mental load of subjects to 41, and the data distribution showed a high degree of consistency.

According to the above qualitative data analysis, VGAGN is not only an algorithmic plug-in which enhances the image generation quality; it from the basic level changes the interaction mode that human and computer create together. The input of CV visual guiding characteristics successfully lets designers' cognitive resources get rid of heavy "machine instruction guessing," permits them to pay attention again to high-level artistic ideas and style exploration, hence accomplishing a jump from "passive creation" to "controllable generation."

3.3 Discussion: Synergy and Controllability in Human-AI Co-creation

After having objective verification on the algorithm's performance and user experience, this section will reply to a deeper theoretical question: In the human-AI co-creation paradigm, is the strong intervention of CV technology able to restrict the artist's freedom, or thus foster a deeper synergistic effect? What are the boundary ranges of this algorithm that is based on rigid features in complex artistic expression activities? Figure 5 exposes the movement path of cognitive resource re-allocation in the workflow of the designer, and also the boundary of the system's ability under the extreme abstract styles.



(a) Cognitive Effort Reallocation **(b) Performance Decay vs. Abstraction**

Figure 5: Analysis of Human-AI synergy and systemic limitations.

As shown in Figure 5(a) of the cognitive resource allocation curve, when using the traditional text-driven model, as the design process progresses, designers spend up to 68% of their energy on "repetitive rendering" (i.e., attempting to eliminate random topological errors by constantly modifying the prompt), severely squeezing the time allocated to high-level "conceptual ideation" to 22%. However, after introducing the VGAGN framework, this distribution undergoes a structural reversal: because CV features precisely take over the generation of the underlying physical topology, the resource allocation for rendering debugging

is significantly reduced to 15%, while the cognitive margin for core concept ideation soars and stabilizes at 65%. This data profoundly demonstrates that CV-based deterministic control does not deprive artists of their creativity. On the contrary, as a powerful "cognitive outsourcing" mechanism, it releases working memory occupied by inefficient trial and error. Designers are able to extricate themselves from the quagmire of "fighting against machine randomness" and refocus their attention on truly irreplaceable high-level aesthetic activities such as color psychology and spatial narrative. However, this collaborative mechanism based on explicit geometry guidance is not without limitations. As shown in the nonlinear performance degradation curve in Figure 5(b), the system exposes clear engineering constraints when faced with extreme abstract styles. When the "Abstraction Index" (quantified by non-Euclidean geometric deformation) of the input hand-drawn sketch exceeds 0.75, the FID quality of the generated image deteriorates sharply, jumping from a stable 15.6 to 34.2, while the local artifact collapse rate reaches as high as 41%. The core reason for this out-of-control phenomenon is that the underlying Canny and MiDaS prior operators are pre-trained based on real-world physical laws (Euclidean geometry and natural perspective). When designers input displaced spaces or minimalist abstract lines similar to Cubism, the rigid $\mathcal{L}_{\text{struct}}$structural constraint loss function becomes an "algorithmic penalty." The model is forced to fit a physical depth that does not exist at all, thus causing serious semantic conflicts and image tearing. In conclusion, CV technology has indeed given generative AI the ability to leap from a "random lottery machine" to a "precise design tool," achieving a deep synergistic effect. However, future research urgently needs to explore adaptive feature constraint thresholds so that algorithms can intelligently identify the abstract intent of sketches, thereby achieving a smooth switch between "rigid physical alignment" and "abstract artistic divergence".

4 Conclusion

To address the spatial loss of control and intent drift challenges faced by generative large models in professional creative industries, this study breaks away from the constraints of a single text-driven approach and constructs a visually guided adaptive generative network (VGAGN).

(1) In terms of data foundation and object organization, the study breaks through the semantic sparsity limitation of the general domain and establishes the “Design-Triplets-52K” high-precision matching dataset that spans two major scenarios: architecture and industry. By innovatively embedding a random space degradation mask at the data ingestion end, the system effectively converges the interaction domain offset error between the perfect algorithm boundary and the real human hand-drawn sketch.

(2) At the algorithm mechanism and verification level, this method completely rewrites the non-convex optimization trajectory of the traditional diffusion model through spatial decoupling of multimodal features and pixel-level structural consistency intervention. Quantitative tests show that the framework not only approaches extremely high fidelity on FID (15.61), but also $\lambda \approx 0.48$achieves an excellent structural alignment rate of 0.89 SSIM on the optimal policy surface. With an inference latency of less than 4 seconds per graph, this study substantially drives the underlying paradigm shift of human-computer co-creation from "black box lottery" to "precise construction".

(3) However, due to the Euclidean geometric prior of the feature extraction operator, the strong physical constraints can easily trigger the semantic collapse of the image when the system is dealing with extremely abstract art such as Cubism or Minimalism. Future research will focus on developing an adaptive visual relaxation operator based on intent recognition, in order to achieve a deeper intelligent fusion between physical rigid alignment and artistic free divergence.

About the Author

Yi Gong was born in 2000. She obtained her master's degree from Taiyuan University of Technology, China. She is currently pursuing a PhD in Design at the School of New Media Art and Design, Beihang University. Her main research interest is Information Art Design.

Shuai Li was born in 1981. He obtained his Ph.D. degree from Beihang University, China. He is currently a professor at the School of Computer Science and Engineering, Beihang University. His main research interests include computer graphics, pattern recognition, computer vision, and medical image processing.

References

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

[2] Peebles, W., Xie, S. (2023). Scalable diffusion models with transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4195-4205.

[3] Qin, Z. (2024). A multimodal diffusion-based interior design AI with ControlNet. Journal of Artificial Intelligence Practice, 7(1), 162-165.

[4] Kumari, N., Zhang, B., Zhang, R., et al. (2023). Multi-concept customization of text-to-image diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1931-1941.

[5] Ruiz, N., Li, Y., Jampani, V., et al. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22500-22510.

[6] Zhang, Y., Dong, W., Tang, F., et al. (2023). Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. ACM Transactions on Graphics (ToG), 42(6), 1-14.

[7] Wang, N., Kim, H., Peng, J., et al. (2025). Exploring creativity in human–AI co-creation: A comparative study across design experience. Frontiers in Computer Science, 7, 1672735.

[8] Herath, S., Bashardoust, A., Bole, Y., et al. (2026). Design principles for text-to-image generative artificial intelligence creativity support tools for visual design. European Journal of Information Systems, 1-26.

[9] Garcia, MB (2025). The paradox of artificial creativity: Challenges and opportunities of generative AI artistry. Creativity Research Journal, 37(4), 755-768.

[10] Oppenlaender, J., Linder, R., Silvennoinen, J. (2025). Prompting AI art: An investigation into the creative skill of prompt engineering. International Journal of Human–Computer Interaction, 41(16), 10207-10229.

[11] Leng, J., Ye, H., Xu, P., et al. (2025). GenFODrawing: Supporting creative found object

drawing with generative AI. IEEE Transactions on Visualization and Computer Graphics.

[12] Turchi, T., Carta, S., Ambrosini, L., et al. (2023). Human-AI co-creation: Evaluating the impact of large-scale text-to-image generative models on the creative process. International Symposium on End User Development, Cham: Springer Nature Switzerland, 35-51.

[13] Mangubat, J. (2026). Preference assessment of generative AI image tools for architectural design generation using MS-TORO framework. Journal of Asian Architecture and Building Engineering, 1-12.

[14] Alamasi, R., Asfour, OS (2026). Applications of generative AI in architectural design education: A systematic review and future insights. Digital, 6(1), 6.

[15] Cao, Q., Zhou, Y. (2025). Research on the application effectiveness of generative AI in design projects from data-driven and sustainable perspectives. Sustainability, 17(23), 10643.

[16] Alharthi, SA (2025). Generative AI in game design: Enhancing creativity or constraining innovation? Journal of Intelligence, 13(6), 60.

[17] Nandakumar, N., Eberhardt, J. (2025). A synthetic image generation pipeline for vision-based AI in industrial applications. Applied Sciences, 15(23), 12600.

[18] Alhabeeb, SK, Al-Shargabi, AA (2024). Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. IEEE Access, 12, 24412-24427.

[19] Bengesi, S., El-Sayed, H., Sarker, MK, et al. (2024). Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Access, 12, 69812-69837.

[20] Bervar, M., Bertoncel, T., Pejić Bach, M. (2026). Generative artificial intelligence and the creative industries: A bibliometric review and research agenda. Systems, 14(2), 138.

[21] Gang, H., Yi, Z. (2026). A study on the convergence and evolution of artificial intelligence and art design from a bibliometric perspective. International Conference on Computer Vision and Digital Art (ICCVDA 2025), SPIE, 14117, 250-257.

[22] Lei, Y., Li, J., Li, Z., et al. (2024). Prompt learning in computer vision: A survey. Frontiers of Information Technology & Electronic Engineering, 25(1), 42-63.

[23] Kim, TS, Ignacio, MJ, Yu, S., et al. (2024). UI/UX for generative AI: Taxonomy, trend, and challenge. IEEE Access, 12, 179891-179911.

[24] Huang, J. (2024). The art of AI: A human-centered AI (HCAI) user study of integrating image-generative tools in visual art workflows: The case of Adobe Firefly.

[25] Jiang, Y., Fan, Y., Liu, Z. (2025). Generative AI in art education: A systematic review of research trends, tool applications, and outcomes (2019–2025). Education Sciences, 16(1), 47.