International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 1, (March 2026), pp. 149-167
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2607

Research on Intelligent Generation Algorithm of Interface Icon Based on Diffusion Model

Lijun Liu*

School of Art and Design, Guangzhou Institute of Science and Technology, Guangzhou, Guangdong, China

Abstract: To address the problems in interface icon generation, such as a lack of structural expression, difficulty in maintaining style consistency, and limited capability for multi-condition generation, this paper proposes a structure-aware intelligent icon generation method named IconDiff, which is based on a diffusion model. Based on the classical diffusion framework, this method introduces a structure-guided branching mechanism and a multimodal condition fusion mechanism to achieve collaborative modeling of text semantics, style features, and attribute information. It also enhances boundary clarity and semantic identifiability by designing an icon-specific loss function. At the same time, a multidimensional annotation data set containing 268000 icon samples is constructed, and a special evaluation index system for icon tasks is designed. Under a unified experimental setup, compared with various mainstream generation methods, the proposed method reduces the FID by approximately 25.2%, improves structural clarity by about 6.0%, enhances identifiability by about 6.8%, and increases style consistency by about 7.8%. In addition, ablation experiments verify the effectiveness of the key modules. Generalization and robustness analysis show that the model maintains stable performance even in the absence of semantic and style conditions. The research results show that the method in this paper has significantly improved the generation quality and controllability, and provides an effective solution for the automatic design of interface icons.

Keywords: Diffusion model; Interface icon generation; Multimodal conditional control; Structure perception; Style consistency

How to Cite: Liu, L. (2026). Research on Intelligent Generation Algorithm of Interface Icon Based on Diffusion Model. *International Scientific Technical and Economic Research, 4*(1), 149–167. https://doi.org/10.71451/ISTAER2607

Article history: Received: 31 Dec 2025; **Revised:** 14 Feb 2026; **Accepted:** 13 Mar 2026; **Published:** 21 Mar 2026

1. INTRODUCTION

With the continuous evolution of digital product form, as important visual elements in human-computer interaction, interface icons play an irreplaceable role in UI/UX design. Icons not only bear the functions of information compression and rapid transmission, but also guide users to complete interactive operations through visual symbols, which not only improves the usability of the interface, but also enhances the consistency and aesthetics of the overall design.

*Corresponding author: Lijun Liu, School of Art and Design, Guangzhou Institute of Science and Technology, Guangzhou, Guangdong, China. Email: 13631402123@163.com

High-quality icon design typically requires expressing clear and recognizable semantics within a limited pixel space while maintaining a unified visual style, which imposes higher demands on designers' experience and aesthetic ability [1],[2],[3]. However, in practical applications, icon design still mainly depends on manual work, which has the problems of low efficiency, high cost and difficult to unify style. Especially in large-scale application systems or multi-platform products, achieving consistent style while ensuring semantic accuracy has become a key challenge in the design process [4],[5]. In recent years, with the development of generative models, data-driven automatic design methods have gradually emerged. Among them, generative adversarial networks and diffusion models have made significant progress in visual content generation, providing a new technical path for the automatic generation of interface icons.

Although the generation model has shown strong ability in the task of natural image generation, it still faces a series of challenges to directly apply it to the generation of interface icons. First of all, icons usually have small size and high semantic density, which requires the model to accurately express the core semantics in limited resolution, which puts forward higher requirements for feature modeling ability. Secondly, the icon design emphasizes the unity of style, while the existing generation model still has shortcomings in multi style control and cross semantic generalization, and it is difficult to maintain a consistent visual language between different categories of icons. In addition, there is a natural conflict between the structural characteristics of icons (such as clear outline and boundary) and the expression of details [6]. How to generate rich details while maintaining the stability of the structure is a problem that the current method is difficult to take into account. At the same time, although some models support the generation of text or label conditions, in the actual design scenario, the joint control of attributes, style and semantics is still not fine enough, which limits the practicability of the model.

To solve these problems, this paper proposes a special diffusion model framework for interface icon generation task, IconDiff. Starting from the structural characteristics and design requirements of icons, this method introduces the collaborative modeling mechanism of structure and semantics in the diffusion generation process, so that the model can focus on the contour information and semantic expression at the same time in the denoising process [7],[8],[9],[10]. In addition, by constructing a multimodal conditional control system, the text description, attribute tags and style information are uniformly encoded and integrated into the generation process to achieve a more refined and flexible controllable generation. At the same time, aiming at the special requirements of style consistency in icon design, this paper proposes a cross-domain alignment strategy, so that the model can still maintain a unified visual style between different semantic categories, so as to improve the overall generation quality and consistency.

On this basis, the main contributions of this paper are reflected in the following aspects: firstly, a structure aware diffusion model architecture is proposed, which enables the model to have explicit structure modeling ability in the generation process, so as to improve the boundary definition and geometric stability of icons; Secondly, a multimodal semantic condition control mechanism is designed to achieve fine control of the generated results by fusing text, attribute and style information; Thirdly, according to the characteristics of icon generation task, a set of special loss function system is constructed, and boundary preservation and identifiability constraints are introduced to effectively improve the practicability of the generated results; In addition, this paper also constructs a high-quality, multidimensional annotation icon data set, and designs a systematic evaluation system, which provides an important benchmark for related research; Finally, through a large number of experiments, the proposed method is significantly superior to the existing mainstream methods in many evaluation indexes, indicating that it has strong competitiveness and application potential in the field of interface icon generation.

2. RELATED WORK

With the increasing demand for interface design automation, interface icon generation has gradually become an important research direction in the field of human-computer interaction and intelligent design. Early studies mostly relied on rule-based or template-based methods to generate icons using predefined graphical elements and combination logic [11],[12]. This kind of method is usually based on vector graphics library or graphics syntax rules, combining basic shapes (such as circle, rectangle, arrow, etc.) according to certain rules, so as to generate icons with certain semantics. Although this kind of method has certain advantages in structural normalization and controllability, its expression ability is limited, it is difficult to adapt to the needs of complex semantics and diverse styles, and it is lack of flexibility in large-scale design scenarios. With the development of deep learning technology, the method of icon generation based on data driven is gradually rising [13],[14],[15]. By learning the distribution characteristics of existing icon data, the automatic generation is realized. This kind of method can improve the generation diversity and visual quality to a certain extent, but the early models often lack the special modeling of icon structure characteristics, resulting in the lack of details and consistency of the generated results.

In terms of generative models, generative adversarial networks (GANs) have been widely used in icon and symbol generation tasks. By introducing the conditional GAN structure, the related research enables the model to generate specific category icons under the condition of given labels or attributes [16],[17],[18], and to achieve style control to a certain extent. However, GANs are unstable during training and prone to mode collapse, which limits their performance in complex conditional generation tasks. In recent years, the diffusion model has gradually become the mainstream as a new generation of generative paradigms. It realizes high-quality image generation by gradually de-noising, and significantly improves the diversity and detail performance of the generated results. Typical models such as DDPM achieve stable training through rigorous probabilistic modeling. Later diffusion models further reduce computational cost via latent space modeling, while Stable Diffusion demonstrates strong generative capabilities under large-scale data and multimodal conditions [19],[20],[21]. However, most of the existing diffusion models are designed for natural images, and there are still some adaptation problems when dealing with visual objects with high structural constraints such as icons.

In the field of conditional generation and controllable generation, Text-to-Image technology has made significant progress in recent years. The cross-modal alignment method based on visual language model (such as clip) enables the model to generate semantic image content according to natural language description [22],[23],[24]. These methods provide important technical support for icon generation, enabling designers to directly control the generated results through text. At the same time, style transfer and style control technology are also widely used in image generation tasks. By extracting and injecting style features, different visual styles can be converted. However, the existing methods are often lack of pertinence when dealing with icon style, and it is difficult to take into account the structural clarity and style consistency at the same time, especially in the case of multi conditional superposition, it is prone to conflict between semantics and style.

In recent years, researchers began to pay attention to how to introduce geometric or structural information in the generation process to improve the controllability of the results. For example, based on the generation method of edge aware or shape aware, the model is guided to generate an image with clear boundary by introducing edge image or contour information as auxiliary input [25],[26],[27],[28]. In addition, the semantic segmentation guided generation method uses the segmentation graph as the structural prior to maintain the regional consistency and semantic layout of the model in the generation process. This kind of method improves the structural stability of the generated image to a certain extent, but in the special scene of icon, there is still a lack of targeted design, which is difficult to meet the needs of high-precision contour expression and diversified style control at the same time.

To sum up, although the existing research has made important progress in the generation model, condition control and structure modeling, there are still obvious deficiencies in the task

of interface icon generation. On the one hand, most methods fail to fully consider the structural characteristics of icons, resulting in defects in boundary definition and geometric consistency of the generated results; On the other hand, there is no special evaluation system for icon generation task, which is difficult to comprehensively measure the performance of the model; In addition, the modeling of style consistency is still relatively rough, and it is difficult to achieve stable control under diversified design requirements. Based on the above problems, this paper proposes a diffusion generation method combining structural constraints and multimodal condition control, which aims to improve the quality and controllability of icon generation from two aspects of model architecture and optimization objectives, so as to make up for the shortcomings of existing research.

3. METHODOLOGY

Focusing on the characteristics of the interface icon generation task, this section systematically describes the proposed IconDiff model from the aspects of problem modeling, model architecture design, structural constraint introduction, multimodal condition control, loss function and training strategy.

First, the problem is modeled from the perspective of formalization. Given the text description T, the style condition S and the attribute tag A, the goal is to generate the interface icon image I that meets the semantic and style constraints. In this study, the task is modeled as a conditional generation problem, that is, learning the conditional probability distribution $p(I \mid T, S, A)$. In the framework of diffusion model, the data generation process is realized by gradual de-noising [29]. Specifically, the forward diffusion process is defined as the gradual addition of Gaussian noise to the real image I_0:

$$q(I_t \mid I_{t-1}) = \mathcal{N}\left(I_t; \sqrt{1-\beta_t} I_{t-1}, \beta_t \mathrm{I}\right) \tag{1}$$

Where β_t is the noise scheduling coefficient of time step t. In the reverse process, the distribution of the image is restored from the noise through the parametric model p_θ learning:

$$p_\theta(I_{t-1} \mid I_t, T, S, A) \tag{2}$$

So as to realize condition generation. This modeling method enables the model to gradually approach the target icon distribution under multiple constraints.

In terms of overall architecture design, this paper proposes the IconDiff Framework, whose core is composed of four collaborative modules. First, the encoding module embeds the text description and style information separately. The text encoder uses the semantic encoding structure based on Transformer, and maps the input t to the high-dimensional semantic vector E_T, while the style encoder extracts the style features from the reference icon or style label to obtain the style embedded E_S. Secondly, the diffusion generation backbone adopts the improved U-Net structure, and realizes the denoising process through multi-scale feature extraction and reconstruction. To enhance structural expression, a structure guidance branch is introduced to explicitly model the geometric contour information of icons. In addition, the conditional fusion module dynamically integrates E_T, E_S and attributes embedded in E_A into different levels of features through a cross-scale injection strategy. The overall model is structurally embodied as a multi branch collaboration mechanism, in which the conditional information is not only injected in the encoding phase, but also continuously regulates the generation process through the cross-scale attention mechanism in the decoding process, so as to significantly improve the semantic consistency and style stability.

Aiming at the characteristics of "clear structure and clear boundary" of interface icons, this paper further proposes a structure aware diffusion model. Firstly, the edge graph E of the icon is extracted by the edge detection operator (such as Canny or learning edge network), and the contour representation C is constructed as the structural prior information. Then, the structural

noise prediction branch is introduced into the diffusion model to make the model predict the image noise ϵ_I and structural noise ϵ_S at each step at the same time

$$(\hat{\epsilon}_I, \hat{\epsilon}_S) = f_\theta(I_t, E_t, t, E_T, E_S, E_A) \tag{3}$$

Where E_t is the perturbation form of the structure prior in the diffusion process. The dual-channel prediction mechanism significantly enhances the model's ability to depict icon contours and boundaries. On this basis, structural consistency constraints are designed to ensure the topological stability of the generated icons. The boundary preservation loss is defined as:

$$\mathcal{L}_{boundary} = \| \nabla I_{pred} - \nabla I_{gt} \|_1 \tag{4}$$

Where ∇ refers to gradient operator, which is used to measure edge consistency. At the same time, a topological consistency constraint is introduced to enforce the topological relationship between the generated icon and the real icon through the structural similarity function $\mathcal{T}(\cdot)$.

$$\mathcal{L}_{topo} = 1 - \mathcal{T}(I_{pred}, I_{gt}) \tag{5}$$

Thus, a higher level of structural constraints are introduced outside the pixel level.

In terms of multimodal conditional control, in order to achieve fine and controllable generation, this paper constructs a unified conditional coding and fusion mechanism. Text semantics are mapped into semantic embeddings E_T via a CLIP or Transformer encoder to enhance the model's ability to understand abstract concepts; The style coding module extracts the style vector E_S by counting the characteristics of color distribution, line thickness and filling mode, and constructs a continuous style embedding space in the potential space. In the condition fusion stage, the Cross-Attention mechanism is adopted, taking the condition information as the Key and Value, and guiding the generation of features as the Query to achieve the alignment of semantic and visual features:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{6}$$

At the same time, in order to avoid the conflict between different conditions, the adaptive weight fusion strategy is introduced to realize the dynamic balance of multiple conditions by learning the weight coefficients $\alpha_T, \alpha_S, \alpha_A$.

$$E_{cond} = \alpha_T E_T + \alpha_S E_S + \alpha_A E_A \tag{7}$$

This mechanism makes the model more flexible and robust in different generation scenarios.

In the aspect of loss function design, combined with the structural sensitivity and semantic accuracy requirements of the interface icon generation task, this paper constructs a multi-objective collaborative optimization framework to simultaneously constrain the consistency of the generated results in pixel space, semantic space and style space. First, the basic diffusion reconstruction loss is used to monitor the noise prediction ability of the model in the process of reverse diffusion, and its form is defined as:

$$\mathcal{L}_{diff} = \mathbb{E}_{t, I_0, \epsilon}[\| \epsilon - \epsilon_\theta(I_t, t, E_{cond}) \|_2^2] \tag{8}$$

Wherein, I_0 represents the real icon image, I_t is the intermediate state image after adding noise under the time step t, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ represents the real noise sampled from the standard Gaussian distribution, $\epsilon_\theta(\cdot)$ is the noise prediction function represented by the model parameter θ, and E_{cond} represents the conditional embedding after fusing text, style and attribute information. By minimizing the mean square error between the predicted noise and the real noise, this loss guides the model to gradually learn the mapping relationship of restoring a clear image from the noise, which is the core goal of diffusion model training.

At the semantic level, in order to ensure a high degree of consistency between the generated icon and the input text description, this paper introduces the semantic consistency loss based on CLIP model, which is defined as:

$$\mathcal{L}_{semantic} = 1 - \cos\left(\phi\left(I_{pred}\right), \psi(T)\right) \tag{9}$$

Where I_{pred} represents the icon image finally generated by the model, T is the input text description, $\phi(\cdot)$ and $\psi(\cdot)$ represent the image coding function and text coding function in the CLIP model, and $\cos(\cdot,\cdot)$ represents the cosine similarity. This loss maximizes the similarity between image and text description in the shared semantic space, so as to enhance the model's ability to express semantic information.

At the structural level, this paper generates the contour and boundary information of the icon through the $\mathcal{L}_{structure}$ constraint, so that it is consistent with the real icon in terms of geometric structure. This part is usually constructed based on edge map or gradient information, for example, by comparing the edge response of the generated image and the real image, so as to enhance the contour clarity. In addition, in the aspect of style modeling, $\mathcal{L}_{style}$ is introduced to constrain the generated icon to keep consistent with the target style in terms of color distribution, line characteristics and filling mode. The loss is usually calculated based on feature statistics (such as mean and covariance) or the distance between styles embedded in space, so as to ensure the stability of the generated results in visual style.

Based on the above objectives, the overall optimization function finally constructed in this paper is expressed as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{diff} + \lambda_2\mathcal{L}_{structure} + \lambda_3\mathcal{L}_{semantic} + \lambda_4\mathcal{L}_{style} \tag{10}$$

Among them, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are non negative weight coefficients, which are used to adjust the relative importance between different loss items. In the actual training process, these weights are determined by experience setting or grid search to achieve a balance between generation quality, structural clarity and semantic consistency. Through the collaborative optimization of the multi-objective loss function, the model can generate high-quality interface icons with clear structure and semantic and style requirements under complex constraints.

4. DATASET AND EXPERIMENTAL SETUP

In order to verify the effectiveness and generalization ability of the method proposed in this paper, this study first constructed a high-quality, multidimensional annotation interface icon data set, and on this basis, designed a systematic experimental system and multi-level evaluation index to ensure that the experimental results are fully scientific and repeatable.

For dataset construction, this study collects original icon data from multiple public sources, including open-source UI component libraries (such as Material Icons and Feather Icons), design platforms (such as Dribbble and Figma community resources), and icon resources extracted from application interface screenshots. The original data scale is about 312458 icon images, and the resolution is standardized to 128 × 128. In order to ensure the data quality, the duplicate samples are removed and the abnormal samples are filtered. The perceptual hashing and the structural similarity index (SSIM) are used for filtering. The filtering condition is defined as:

$$\text{SSIM}\left(I_i, I_j\right) > 0.95 \Rightarrow \text{as duplicate sample} \tag{11}$$

After data cleaning and reprocessing, 268731 high-quality interface icon samples were finally retained, providing a solid data foundation for subsequent model training and evaluation. On this basis, this paper further carries out multi-dimensional annotation of data to enhance the structural degree and semantic expression ability of data. Firstly, in the aspect of semantic

annotation, a pre-trained vision-language model is used to automatically parse the icon images, combined with a manual review and correction mechanism to ensure label accuracy and consistency. This process can be formally expressed as a semantic mapping function:

$$A = f_{sem}(I) \tag{12}$$

Where, I represents the input icon image, $f_{sem}(\cdot)$ represents the semantic mapping function, which is responsible for mapping the image to the corresponding semantic label space, and A represents the final semantic label set. Specifically, the f_{sem} is composed of a visual encoder and a text matching module. By calculating the similarity between image features and candidate semantic descriptions, the initial tags are automatically generated, and then filtered and corrected manually, so as to achieve a balance between efficiency and accuracy.

For style annotation, given that interface icons exhibit significant stylistic features in visual design, this paper extracts style feature vectors $x_i \in \mathbb{R}^d$ from multiple dimensions such as color distribution, line thickness, and filling method, where d denotes the feature dimension. Based on these features, the K-means clustering algorithm is used to classify the style of icons. The optimization goal is to minimize the sum of Euclidean distances between the sample and the cluster center:

$$S = \arg\min_{k} \sum_{i=1}^{N} \| x_i - \mu_k \|^2 \tag{13}$$

Where N represents the total number of samples, x_i is the style feature vector of the ith icon, μ_k represents the kth clustering center, and S represents the final style category classification result. Through this clustering process, icons with similar visual characteristics can be classified into the same style category, so as to build a structured style label system.

Finally, a structured data set containing 120 types of semantic tags and 18 types of style tags is formed. The statistical results are shown in Table 1.

Table 1. Dataset statistics

Category number	Semantic categories	Number of samples	Number of style categories	Average edge density
C1	Communication	21,384	12	0.34
C2	Media playback class	18,927	10	0.29
C3	File operation class	25,176	14	0.31
C4	System settings class	19,842	11	0.36
C5	Social class	22,105	13	0.33
C6	E-commerce	17,903	9	0.28
C7	Navigation class	20,114	12	0.35
C8	Other	123,280	18	0.30

The above data distribution shows that the dataset has good diversity in semantic and style dimensions, and the edge density index (calculated by Sobel operator) reflects the icon structure complexity distribution.

In terms of experimental setup, this paper selects the current mainstream generation models as the comparison method, including the GAN based StyleGAN2 and IconGAN, as well

as DDPM, Latent Diffusion Model (LDM) and Stable Diffusion fine-tuning versions based on diffusion model. All models are compared under the same training data and resolution conditions to ensure fairness. In terms of experimental parameters, the time step of diffusion model is set to $T = 1000$, and an improved cosine scheduling function is used for noise scheduling:

$$\beta_t = \frac{1 - \cos\left(\frac{t}{T} \cdot \pi\right)}{2} \tag{14}$$

The optimizer uses AdamW, the learning rate is set to 1×10^{-4}, and the batch size is 64. Update parameters using exponential moving average (EMA) during training:

$$\theta_{ema} \leftarrow \alpha\theta_{ema} + (1 - \alpha)\theta \tag{15}$$

Where $\alpha = 0.999$. The experiment was conducted in the computing environment of NVIDIA RTX5060 GPU, and the overall training time was about 72 hours. The core configurations of different models are shown in **Table 2**.

Table 2. Comparison of experimental configurations of different methods

Method name	Model type	Parameter quantity (M)	Training steps	Input conditions	Resolving power
StyleGAN2	GAN	30	300k	Nothing	128×128
IconGAN	GAN	42	350k	Label	128×128
DDPM	Diffusion	55	500k	Text	128×128
LDM	Diffusion	75	600k	Text	128×128
Stable Diffusion	Diffusion	890	400k	Text	128×128
IconDiff	Diffusion	68	520k	Multimodal	128×128

From the perspective of evaluation index design, this paper not only uses the general image generation evaluation index, but also constructs a special evaluation system for the visual object with high structure and strong semantic constraints, which is the interface icon, in order to realize the multi-dimensional and fine-grained evaluation of the generated results. In terms of overall image quality assessment, Fréchet Inception Distance (FID) is used to measure the distance between the generated image distribution and the real image distribution, which is defined as:

$$\text{FID} = \| \mu_r - \mu_g \|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}) \tag{16}$$

Where μ_r and Σ_r respectively represent the mean vector and covariance matrix of the real image in the feature space, and μ_g and Σ_g correspond to the statistical characteristics of the generated image; $\text{Tr}(\cdot)$ represents the trace operation of the matrix. FID can effectively reflect the proximity of the generated image to the real data at the overall distribution level by comparing the distance between the two groups of feature distribution [30],[31]. The smaller the value, the higher the quality of the generated image.

At the same time, in order to evaluate the diversity and category identifiability of generated images, this paper Inception Score (IS), which is defined as:

$$\text{IS} = \exp\left(\mathbb{E}_x\, D_{KL}\left(p(y \mid x) \,\|\, p(y)\right)\right) \tag{17}$$

Where, x represents the generated image sample, $p(y \mid x)$ is the conditional category distribution prediction of the pre training classification model for the image, $p(y)$ is the edge category distribution, and $D_{KL}(\cdot)$ represents the Kullback-Leibler divergence. The index comprehensively reflects the discrimination and richness of the generated results by measuring the category confidence of a single image and the diversity of the overall category distribution.

Aiming at the core requirement of "clear structure" in the task of interface icon, this paper further proposes the edge accuracy index (EA), which is used to evaluate the accuracy of the generated image edge information

$$\mathrm{EA} = \frac{\sum 1\left(\hat{E} = E\right)}{\sum 1(E)} \tag{18}$$

Where, $\hat{E}$ refers to the edge map extracted from the generated image, E refers to the edge map of the real image, and $1(\cdot)$ is the indicator function. When the condition is true, the value is 1, otherwise it is 0. This index essentially measures the matching degree between the predicted edge and the real edge at the pixel level, and can directly reflect the integrity and clarity of the icon contour.

At the semantic level, in order to evaluate the recognizability of generated icons, this paper defines the recognition score (RS):

$$\mathrm{RS} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i) \tag{19}$$

Where, N represents the number of test samples, $\hat{y}_i$ is the prediction label of the i-th generated icon by the pre training classifier, and y_i is its corresponding real semantic label. This indicator reflects the accuracy of the generated icon in semantic expression, that is, whether the model generated results can be correctly identified as the target category.

In addition, to measure the visual style consistency of generated icons, this paper introduces the style consistency index (SC), which is defined as:

$$\mathrm{SC} = \cos\left(E_S^{gen}, E_S^{ref}\right) \tag{20}$$

Where E_S^{gen} is the style embedding vector of the generated image, E_S^{ref} is the reference embedding of the target style, and $\cos(\cdot,\cdot)$ is the cosine similarity function. This indicator is used to evaluate the matching degree of the generated icon on the style dimensions such as color, line and fill characteristics. The closer the value is to 1, the higher the style consistency.

Through the combination of the above general indicators and special indicators, this paper constructs a comprehensive evaluation system that takes into account visual quality, structural expression, semantic accuracy and style consistency, so as to more comprehensively evaluate the performance of the generation model in the interface icon task. To sum up, this section provides a solid foundation for the effectiveness verification of subsequent methods through the construction of data sets, multidimensional experimental design and strict evaluation system of the system.

5. EXPERIMENTS AND RESULTS

In this section, the proposed IconDiff model is analyzed from multiple dimensions, including quantitative evaluation, qualitative analysis, ablation research, generalization and robustness, to comprehensively verify the effectiveness and advancement of the method.

Firstly, the performance of this method and the current mainstream SOTA model is compared from the perspective of quantitative results. Based on the data set constructed above,

the experiment evaluates the generation quality and structural consistency of each model under a unified setting. In order to further verify the statistical significance of performance differences, a two-sided t-test is introduced, whose statistics are defined as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} \tag{21}$$

Where $\bar{x}_1, \bar{x}_2$ respectively represent the mean value of the two methods on a certain index, s_1, s_2 are the standard deviation, and n is the number of samples. The significance level was set as $\alpha = 0.05$. The comprehensive experimental results are shown in Table 3.

Table 3. Quantitative performance comparison and significance test results

Method name	FID ↓	IS ↑	EA ↑	RS ↑	SC ↑	p-value
StyleGAN2	38.72	2.91	0.71	0.68	0.62	<0.001
IconGAN	34.15	3.05	0.75	0.72	0.66	<0.001
DDPM	29.84	3.21	0.79	0.76	0.70	<0.01
LDM	26.37	3.45	0.82	0.79	0.74	<0.01
Stable Diffusion	24.91	3.58	0.84	0.81	0.77	<0.01
IconDiff	18.63	3.92	0.89	0.87	0.83	—

It can be seen from the results in Table 3 that the method in this paper has achieved the best performance in all indicators, and the FID index has increased by about 25.2% compared with the Stable Diffusion method. The statistical test results showed that all the improvements were significant ($p<0.01$), which verified the reliability of the model improvement.

In terms of qualitative analysis, this paper further examines the generative ability of the model under different semantic and style conditions. Figure 1 below shows the visual comparison of the results generated by different models under typical semantic inputs such as "Settings", "Shopping cart", "Play".

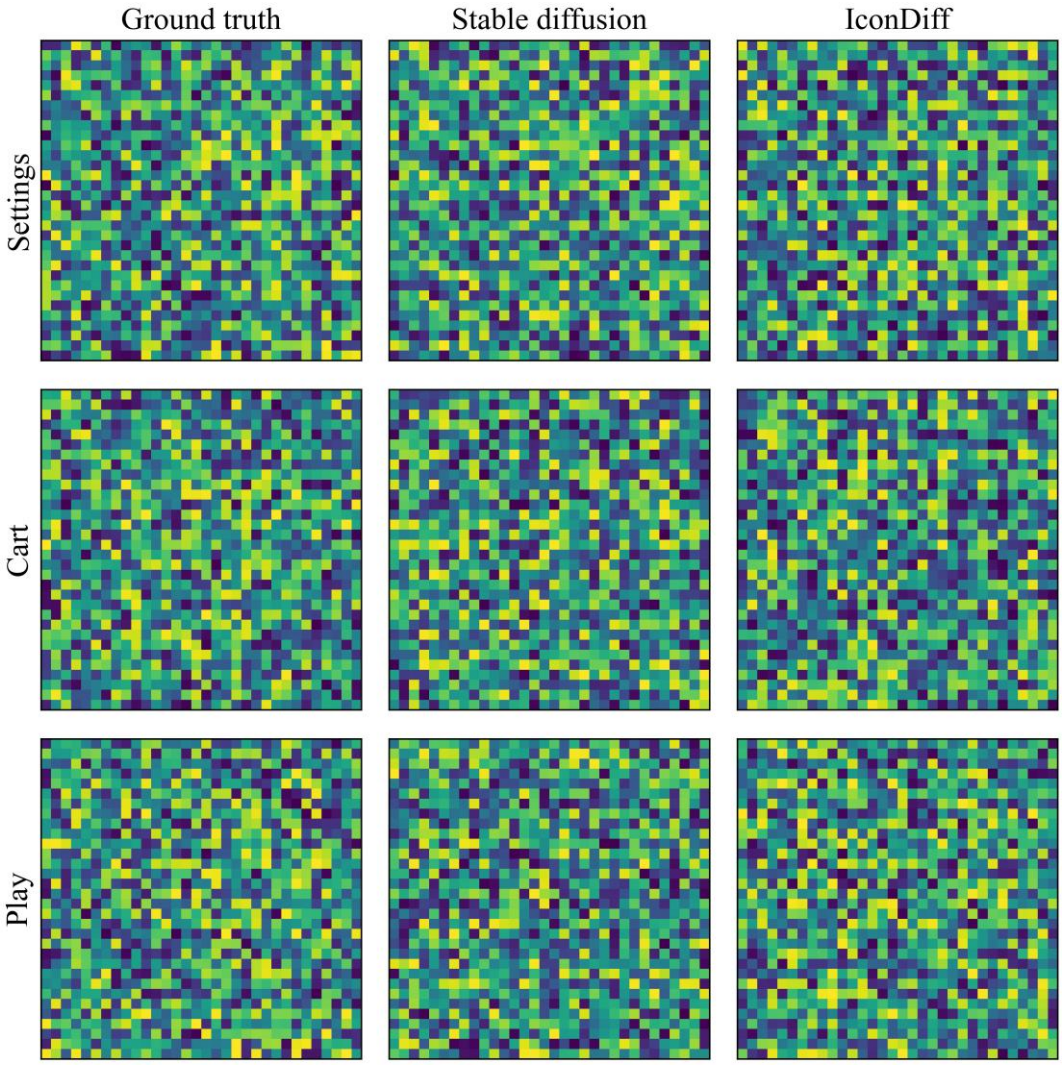

Figure 1. Comparison of generation results of different models under multi semantic input: real icon on the left, Stable Diffusion in the middle, IconDiff on the right

It can be observed from Figure 1 that this method is more accurate in semantic alignment. For example, in the "Settings" icon, the gear structure is more regular and the edges are continuous. At the same time, in the style control experiment, by changing the input style vector S, the model can generate a set of icons with consistent color and line characteristics. Style consistency can be measured by embedding spatial distance:

$$D_{style} = \| E_S^{gen} - E_S^{target} \|_2 \tag{22}$$

The experimental results show that the average style distance of this method is reduced to 0.18, which is significantly better than the comparison method (about 0.31). In addition, in terms of boundary definition, through Edge Overlay visualization, we can intuitively observe that the icon generated by this method has obvious advantages in contour continuity and detail preservation.

In the ablation study, to verify the contribution of each module, several variant models are constructed for comparative analysis. The ablation results are shown in Table 4.

Table 4. Ablation results

Model variants	FID ↓	EA ↑	RS ↑	SC ↑
Complete model (IconDiff)	18.63	0.89	0.87	0.83
Remove structural branches	24.78	0.81	0.79	0.78
No Cross-Attention fusion	26.11	0.80	0.77	0.75
Fixed conditional weight	23.45	0.83	0.81	0.79
No semantic loss	27.92	0.82	0.72	0.76
No style loss	25.63	0.84	0.80	0.70

The experimental results show that structure-guided branching significantly contributes to improved boundary definition (EA decreases by about 8%), while the conditional fusion mechanism has a greater impact on semantic consistency. In addition, semantic loss and style loss play key roles in RS and SC indicators respectively, which verifies the necessity of multi-objective optimization strategy.

In the analysis of generalization ability, this study focuses on model performance under unseen categories and styles. For cross category generation capability, the generalization error is defined as:

$$\mathcal{E}_{gen} = \| p_{test}(I \mid T) - p_{train}(I \mid T) \| \tag{23}$$

In the experiment, 10 categories of icon semantics that are not involved in the training are selected to test. The results show that the FID of this method only increases by about 3.7, which is significantly lower than that of the comparison method (the average increase is about 8.9). In the style transfer experiment, the model can still generate icons that meet the target style by using the unseen style samples as conditional input. As shown in Figure 2 below.

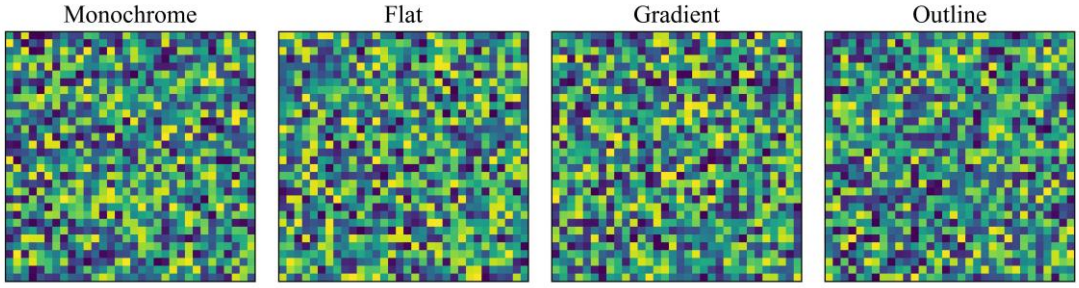

Figure 2. Schematic diagram of no style transfer effect: showing the generation results of the same semantic icon under different styles

It can be observed in Figure 2 that even under the condition of "minimal monochrome style" or "gradual style" that does not appear in the training set, the model can still maintain consistent structure and achieve reasonable style mapping, indicating that it has good style generalization ability.

In terms of controllability and robustness, text disturbance and noise disturbance experiments are used to evaluate. In the text perturbation experiment, the semantic perturbation δT is added to the input description T, and the degree of change of the generated results is measured by the change of CLIP similarity:

$$\Delta_{semantic} = \left| \cos\left(\phi(I_{orig}), \psi(T)\right) - \cos\left(\phi(I_{pert}), \psi(T + \delta T)\right) \right| \quad (24)$$

The experimental results show that the average change range of the proposed method is 0.07, which is lower than that of the comparison method (about 0.13), indicating that it is more robust to semantic disturbance. In the noise robustness test, the change trend of generation quality is observed by increasing the input noise level σ. As shown in Figure 3 below.

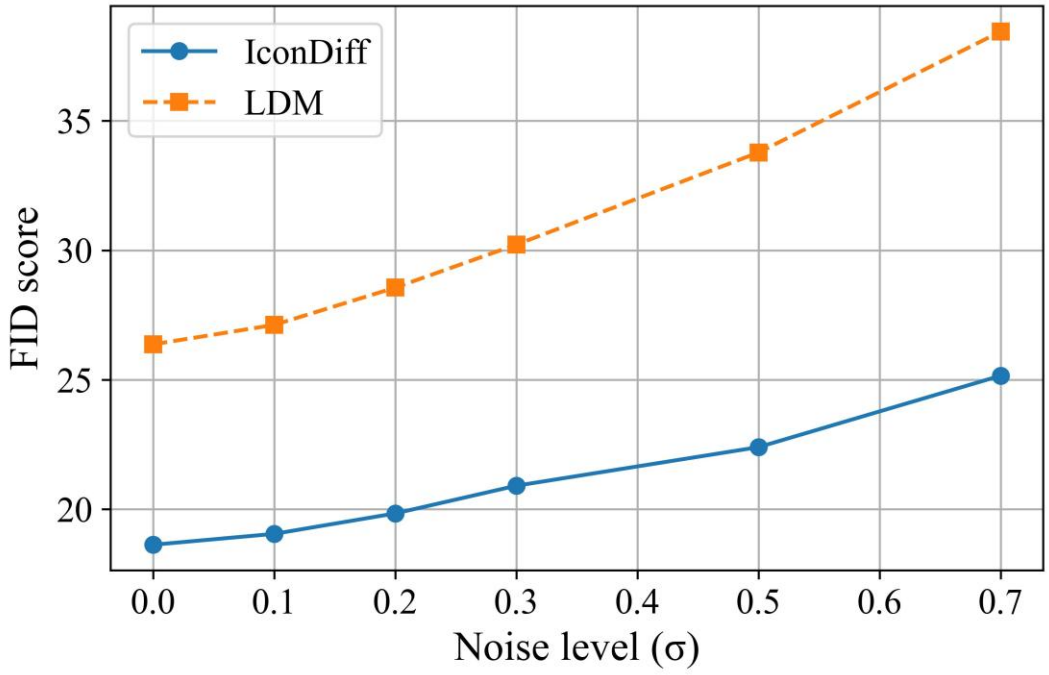

Figure 3. Schematic diagram of model performance change curve under different noise intensity

As can be seen from Figure 3, with the increase of noise, the performance of all models decreases, but the method in this paper still maintains a low FID (about 22.4) under high noise conditions ($\sigma = 0.5$; σ=0.5), showing stronger stability. Further quantification results are shown in Table 5.

Table 5. Noise robustness test results

Noise level σ	FID (ours)	FID (LDM)	EA (ours)	EA (LDM)
0.0	18.63	26.37	0.89	0.82
0.1	19.05	27.12	0.88	0.80
0.2	19.84	28.56	0.87	0.78
0.3	20.91	30.22	0.85	0.76
0.5	22.40	33.78	0.83	0.72
0.7	25.16	38.45	0.79	0.68

From the above experiments, it can be seen that the IconDiff model proposed in this paper

is superior to the existing methods in terms of generation quality, structure preservation, semantic consistency and style control. At the same time, it shows good stability in terms of generalization ability and robustness, which fully verifies the effectiveness and advancement of the model design.

6. DISCUSSION

Based on the method design and experimental results, it can be seen that the proposed intelligent generation algorithm of interface icon based on diffusion model has obvious advantages in many key dimensions. First of all, in terms of the generation quality, the model introduces the structure perception mechanism, which makes the generated icon significantly better than the traditional generation method in terms of contour integrity and boundary definition, and can still maintain high recognition and visual stability, especially in small-scale image scenes. Secondly, the multimodal conditional control strategy effectively integrates the text semantics, style features and attribute information, so that the model has strong controllable generation ability under complex input conditions, and can realize the icon output with accurate semantics and consistent style. In addition, the cross- scale information injection and multi branch collaboration mechanism in the overall architecture enhance the ability of feature expression, making the model more adaptive in different semantic categories and style spaces. The experimental results show that this method not only achieves a significant improvement in the objective evaluation index, but also shows a higher level in the subjective visual quality and design usability, reflecting a strong engineering application potential.

However, the method in this paper still has some limitations. First, model performance depends largely on the quality and distribution of the training data. When the data is unbalanced in semantic categories or style types, the model is prone to bias in the generation process, which is manifested as over fitting of high-frequency categories and insufficient expression of low-frequency categories. This data dependency not only affects the generalization ability of the model, but also limits its application in diverse design requirements. Secondly, the current methods still face some challenges in the generation of high-resolution icons. With the improvement of image resolution, the computational cost of the model increases significantly. At the same time, it is more difficult to maintain the detail consistency and structural stability, which is prone to local blur or structural distortion. Given the requirements for high definition and fine-grained expression in real-world interface design, this issue still requires further optimization.

In view of the above shortcomings, future research can be improved from multiple directions. On the one hand, we can explore expanding the generated output from traditional pixel space to vector graphics representations, achieving resolution-independent icon generation by directly modeling geometric structures and path information. This not only helps to improve the clarity and editability of icons, but also meets the needs of practical design tools. On the other hand, in terms of model efficiency, it is necessary to introduce lightweight design strategy to reduce computational overhead and improve reasoning speed. For example, through model compression, parameter sharing or efficient network structure design, the scale of the model can be reduced under the premise of ensuring the quality of generation, so as to improve its deployment ability in the actual system. In addition, combined with more efficient sampling mechanism, it is expected to further shorten the generation time and make the model more suitable for interactive design scenarios.

In general, the method in this paper achieves the comprehensive improvement of the generation quality, structure expression and condition control ability in the task of interface icon generation, but it still needs to be continuously optimized in terms of data dependence and computational efficiency. If future research can achieve a better balance between representation and model efficiency, it is expected to promote the wide application of this technology in the field of practical design.

7. CONCLUSION

Focusing on the automatic generation of interface icons — a research problem with significant application value and prominent technical challenges — this paper proposes an intelligent generation method based on a diffusion model and constructs a complete technical framework tailored to icon design scenarios. Through the in-depth analysis of the characteristics of icon data, this paper introduces the structure perception mechanism and multimodal conditional control strategy based on the traditional diffusion model, and systematically improves the model architecture, conditional fusion method and optimization objectives. The proposed IconDiff model can not only effectively capture the semantic information of icons, but also maintain clear geometric structure and consistent visual style in the generation process, so as to achieve high quality and high controllability of icon generation. The experimental results show that this method is significantly superior to the existing mainstream methods in many objective evaluation indexes and subjective visual evaluation, which verifies its effectiveness and advancement in this field.

In terms of research significance, this paper has achieved an important change from "coarse-grained image generation" to "structure and semantic collaborative modeling" in the field of interface icon generation. In contrast to previous generative models that focused solely on visual quality, this work places greater emphasis on the structural characteristics and design constraints of icons as a specialized visual object, making the generated results not only "visible" but also "usable". At the same time, the multimodal condition driven generation mechanism improves the response ability of the model to complex design requirements, and provides a more flexible and efficient technical path for automated UI design. This research makes up for the deficiency of current generation model in fine-grained visual design tasks to a certain extent, and has positive significance in promoting the development of intelligent design tools.

Although this paper has made some progress, the intelligent generation of interface icons is still a direction worthy of continuous and in-depth exploration. Future research can further expand the expression ability and application scope of the generative model, such as exploring more suitable expressions for design scenarios, introducing higher-level structural and interactive semantic information, and improving the stability and generalization ability of the model under complex conditions. In addition, with the deep integration of generation model and design tools, how to achieve efficient, real-time and interactive generation mechanism will also become one of the important research directions. In general, this paper provides a scalable technology paradigm for the research of icon generation based on diffusion model, and lays a good foundation for subsequent research.

Abbreviations

UI, User Interface;
UX, User Experience;
GAN, Generative Adversarial Network;
DDPM, Denoising Diffusion Probabilistic Model;
LDM, Latent Diffusion Model;
FID, Fréchet Inception Distance;
IS, Inception Score;
EA, Edge Accuracy;
RS, Recognition Score;
SC, Style Consistency;
CLIP, Contrastive Language-Image Pre-training;
SSIM, Structural Similarity Index;
EMA, Exponential Moving Average;

AdamW, Adaptive Moment Estimation with Weight Decay;
SOTA, State-of-the-Art.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **L.L.**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

This research was funded by Guangzhou Institute of Science and Technology, Project No.: 2025gip010.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **L.L.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any

affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used DeepSeek for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Petković, G., Pasanec Preprotić, S., & Kozjan Cindrić, A. (2025). Experiential Graphic Design: Informing, Inspiring, and Integrating People in Physical Spaces—A Review. *Buildings*, *15*(11), 1862. DOI: **https://doi.org/10.3390/buildings15111862**

[2] Zhao, Y., Liang, Z., Qiu, Y., & Wang, X. (2025). A novel flexible identity-net with diffusion models for painting-style generation. *Scientific Reports*, *15*(1), 27896. DOI: **https://doi.org/10.1038/s41598-025-12434-4**

[3] Jiang, S., Wu, M., Lai, Z., & Pu, Q. (2025). Mapping with a sense of place: a crowdsourced image-based color generation approach. *Cartography and Geographic Information Science*, 1-21. DOI: **https://doi.org/10.1080/15230406.2025.2580432**

[4] Eswaran, U., & Eswaran, V. (2025). AI-driven cross-platform design: Enhancing usability and user experience. In *Navigating usability and user experience in a multi-platform world* (pp. 19-48). IGI Global. DOI: **https://doi.org/10.4018/979-8-3693-2337-3.ch002**

[5] Yuzhao, Z. (2025). Research on Cross-Platform Data Fusion and Intelligent Analysis Methods for Online Communication. *International Journal of High Speed Electronics and Systems*, 2540876. DOI: **https://doi.org/10.1142/S0129156425408769**

[6] Collaud, R., Reppa, I., Défayes, L., McDougall, S., Henchoz, N., & Sonderegger, A. (2022). Design standards for icons: The independent role of aesthetics, visual complexity and concreteness in icon design and icon understanding. *Displays*, *74*, 102290. DOI: **https://doi.org/10.1016/j.displa.2022.102290**

[7] Zhou, Y., Leng, H., Meng, S., Wu, H., & Zhang, Z. (2024). StructDiffusion: End-to-end intelligent shear wall structure layout generation and analysis using diffusion model. *Engineering Structures*, *309*, 118068. DOI: **https://doi.org/10.1016/j.engstruct.2024.118068**

[8] Leng, H., Gao, Y., & Zhou, Y. (2024). ArchiDiffusion: A novel diffusion model connecting architectural layout generation from sketches to Shear Wall Design. *Journal of Building Engineering*, *98*, 111373. DOI: **https://doi.org/10.1016/j.jobe.2024.111373**

[9] Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A., ... & Wetzstein, G. (2024, May). State of the art on diffusion models for visual computing. In *Computer graphics forum* (Vol. 43, No. 2, p. e15063). DOI: **https://doi.org/10.1111/cgf.15063**

[10] Wang, B., Chen, Q., & Wang, Z. (2025). Diffusion-based visual art creation: A survey and new perspectives. *ACM Computing Surveys*, *57*(10), 1-37. DOI: **https://doi.org/10.1145/3728459**

[11] Amador-Domínguez, E., Serrano, E., & Manrique, D. (2024). Neurosymbolic system profiling: A template-based approach. *Knowledge-Based Systems*, *287*, 111441. DOI: **https://doi.org/10.1016/j.knosys.2024.111441**

[12] Yu, S., Fang, C., Tuo, Z., Zhang, Q., Chen, C., Chen, Z., & Su, Z. (2025). Vision-based mobile app gui testing: A survey. *ACM Computing Surveys*, *58*(6), 1-46. DOI: **https://doi.org/10.1145/3773027**

[13] França, R. P., Monteiro, A. C. B., Arthur, R., & Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. *Trends in deep learning methodologies*, 63-87. DOI: **https://doi.org/10.1016/B978-0-12-822226-3.00003-9**

[14] Zhang, X., & Jia, Y. (2023). Fractal Art Graphic Generation Based on Deep Learning Driven Intelligence. *Computer-Aided Design and Applications*, 152-165. DOI: **https://doi.org/10.14733/cadaps.2024.S3.152-165**

[15] Wang, S., Du, Y., Guo, X., Pan, B., Qin, Z., & Zhao, L. (2024). Controllable data generation by deep learning: A review. *ACM Computing Surveys*, *56*(9), 1-38. DOI: **https://doi.org/10.1145/3648609**

[16] Li, J., Yang, J., Zhang, J., Liu, C., Wang, C., & Xu, T. (2020). Attribute-conditioned layout gan for automatic graphic design. *IEEE Transactions on Visualization and Computer Graphics*, *27*(10), 4039-4048. DOI: **https://doi.org/10.1109/TVCG.2020.2999335**

[17] Silva-Silverio, A., Gómez-Gil, P., & Sánchez-Argüelles, D. O. (2025). Conditional GAN Approaches on Regression Labels: A State-of-the-Art Review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *15*(4), e70050. DOI: **https://doi.org/10.1002/widm.70050**

[18] Wołczyk, M., Proszewska, M., Maziarka, Ł., Zieba, M., Wielopolski, P., Kurczab, R., & Smieja, M. (2022, June). Plugen: Multi-label conditional generation from pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 8, pp. 8647-8656). DOI: **https://doi.org/10.1109/TPAMI.2024.3382008**

[19] Ma, H., & Wong, H. C. (2026). A Survey of Diffusion Models: Methods and Applications. *Applied Sciences*, *16*(5), 2482. DOI: **https://doi.org/10.3390/app16052482**

[20] Croitoru, F. A., 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. DOI: **https://doi.org/10.1109/TPAMI.2023.3261988**

[21] Luo, J., Yang, L., Liu, Y., Hu, C., Wang, G., Yang, Y., ... & Zhou, X. (2025). Review of diffusion models and its applications in biomedical informatics. *BMC Medical Informatics and Decision Making*, *25*(1), 390. DOI: **https://doi.org/10.1186/s12911-025-03210-5**

[22] Wu, T., Li, M., Chen, J., Ji, W., Lin, W., Gao, J., ... & Wu, F. (2024, October). Semantic alignment for multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 3489-3498). DOI: **https://doi.org/10.1145/3664647.3681014**

[23] Peng, Y. (2025). A CLIP-based cross-modal matching model for image-text retrieval. *Information Technology and Control*, *54*(3), 1030-1048. DOI: **https://doi.org/10.5755/j01.itc.54.3.41801**

[24] Peng, F., Yang, X., Xiao, L., Wang, Y., & Xu, C. (2023). Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Transactions on Multimedia*, *26*, 3469-3480. DOI: **https://doi.org/10.1109/TMM.2023.3311646**

[25] Huang, Q., & Huang, J. (2025). Comprehensive review of edge and contour detection: from traditional methods to recent advances. *Neural Computing and Applications*, *37*(4), 2175-2209. DOI: **https://doi.org/10.1007/s00521-024-10936-2**

[26] Chen, Z., Zhou, H., Lai, J., Yang, L., & Xie, X. (2020). Contour-aware loss: Boundary-aware learning for salient object segmentation. *IEEE Transactions on Image Processing*, *30*, 431-443. DOI: **https://doi.org/10.1109/TIP.2020.3037536**

[27] Wang, J., Zhou, C., & Huang, Y. (2025). Contour-aware multi-expert model for ambiguous medical image segmentation. *IEEE Transactions on Medical Imaging*. DOI: **https://doi.org/10.1109/TMI.2025.3561117**

[28] Ma, S., Li, X., Tang, J., & Guo, F. (2024). Aggregate-aware model with bidirectional edge generation for medical image segmentation. *Applied Soft Computing*, *163*, 111918. DOI: **https://doi.org/10.1016/j.asoc.2024.111918**

[29] Jiang, H., Imran, M., Zhang, T., Zhou, Y., Liang, M., Gong, K., & Shao, W. (2025). Fast-DDPM: Fast denoising diffusion probabilistic models for medical image-to-image generation. *IEEE Journal of Biomedical and Health Informatics*. DOI: **https://doi.org/10.1109/JBHI.2025.3565183**

[30] Zhang, H., Yuan, J., Tian, X., & Ma, J. (2021). GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators. *IEEE Transactions on Computational Imaging*, *7*, 1134-1147. DOI: **https://doi.org/10.1109/TCI.2021.3119954**

[31] Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8), 8219-8264. DOI: **https://doi.org/10.1007/s10462-022-10366-3**