Artistic Creation Inspiration Generation System Based on Diffusion Model

Jie Zhou[1], Yu Zhou[2], Lu Yang[2]
[1]Department of Oil Painting, School of Fine Arts, Sichuan Fine Arts Institute, Chongqing 401331, China
[2]College of Fine Arts, Chongqing Normal University, Chongqing 401331, China
Corresponding author: Lu Yang (e-mail: yangluart@163.com)

ABSTRACT In traditional artistic creation, inspiration relies heavily on subjective experience and occasional associations, resulting in limited systematicity and controllability during the early conception stage. Similar challenges also exist in intelligent visual design and information presentation for engineering applications, where efficient human–machine interaction and semantic visualization are increasingly important for electromagnetic communication systems and digital design environments. This paper proposes a text-guided diffusion model for artistic inspiration generation, particularly suitable for visual creation tasks such as illustration and conceptual art while remaining applicable to product innovation fields. The proposed framework integrates CLIP semantic embedding with conditional control mechanisms into a customized UNet architecture based on Stable Diffusion, enabling user text prompts to guide image synthesis through multi-layer cross-attention. Classifier-free guidance (CFG) and a multi-scale noise scheduling strategy are further introduced to improve style controllability, detail representation, and generation stability while supporting iterative refinement. Experimental results demonstrate stable performance with high semantic consistency (CLIP similarity of 0.82), effective style control (0.76), and favorable image quality (FID of 18.5). User evaluations achieve scores of 4.33 for creative inspiration and 4.49 for overall satisfaction, confirming that the proposed system provides an efficient and expressive visual inspiration generation tool. The framework also offers a systematic approach for multimodal semantic visualization and intelligent content generation, which may provide useful references for visual information representation and human–machine collaborative design in advanced electromagnetic and communication-related applications.

INDEX TERMS diffusion model, conditional image generation, multimodal semantic representation, electromagnetic information visualization, classifier-free guidance

I. INTRODUCTION

WITH the continuous expansion of artificial intelligence in the field of content generation, visual generation models have gradually penetrated into multiple humanistic expression scenarios such as creative design [1,2] and visual art. The capacity of these models to rapidly create and visualize concepts is being adopted across diverse sectors [3,4]. At present, systematic research on the inspiration generation process is still relatively weak, making it difficult to effectively provide creators with efficient, flexible and controllable image references, which restricts the application and expansion of human-computer collaborative creation in the field of art. In this context, exploring an image generation mechanism that is both controllable, diverse, and semantically consistent to assist artistic conception has important theoretical significance and practical value. The theoretical underpinnings of generating diverse, high-fidelity concepts are transferable to content-heavy industries.

In recent years, diffusion models, as an important direction of generative modeling, have shown advantages in image quality, semantic expression, and style diversity [5,6]. By reconstructing images, they improve the ability to restore high-frequency details and preserve structures, and are suitable for scenes that pursue complex compositions and delicate styles in artistic creation. Related research focuses on the design of the basic framework for text-driven image generation, covering directions such as conditional modeling, semantic guidance, and style transfer [7,8]. However, most existing studies focus on general image generation tasks, and there is still a lack of modeling and system construction for the specific context of "inspiration generation" in artistic creation scenarios.

This paper further constructs an inspiration generation experimental framework for artistic creation, and evaluates the comprehensive performance of the method in terms of semantic consistency, style expression and inspiration value through quantitative indicators and user feedback. The research results verify the applicability and superiority of the proposed method in inspiration generation tasks, and provide a feasible path for building a creative assistance

system based on the diffusion model. The demonstrated capability to provide reliable, data-driven inspiration is highly transferable.

II. RELATED WORKS

In the conception stage of artistic creation, creators often need to stimulate abstract thinking and emotional associations through concrete media [9,10], but traditional ways of obtaining inspiration such as life observation, emotional accumulation and cross-media stimulation [11,12]. These methods have obvious limitations in terms of information sources, associative paths, and forms of expression. In order to enhance the image support in the early stages of conception, some studies have developed reference libraries and style assistance tools [13,14]. However, such tools usually lack semantic understanding and generation capabilities, making it difficult to achieve deep coordination between the creator's intention and the image content, limiting their practicality in the creative divergence stage.

In recent years, diffusion models have made progress in image generation tasks and have outperformed traditional generative models in text-to-image generation tasks [15,16]. Representative models such as DALL·E 2 and Stable Diffusion all introduce text-guided mechanisms based on the diffusion model framework to achieve effective mapping of semantic information to image space [17,18]. Unlike GAN, the diffusion model can more finely control the semantic structure and local details of the image during the generation process, and is suitable for artistic creation tasks that need to express multiple layers of meaning and present abstract visual language. Current research is also gradually expanding in the direction of multi-style fusion and cross-modal hybrid guidance, providing a broader modeling space for creative generation.

III. CONSTRUCTION OF TEXT-DRIVEN DIFFUSION MODEL IN ARTISTIC INSPIRATION GENERATION

SYSTEM FRAMEWORK OVERVIEW

This system is based on a text-guided diffusion generation architecture. The overall process includes five modules: text input, semantic encoding, conditional diffusion generation, multi-scale scheduling, and image output. Users input their creative intentions through natural language, and the text information is extracted and semantically embedded by the Text Encoder (ViT-B/32) of the CLIP model, and adapted to the input requirements of the Cross-Attention module in UNet through linear mapping. These conditional features are inserted in multiple layers in the diffusion network to achieve spatial scale fusion of semantic control. To improve the performance of image details, the system designs a multi-scale noise scheduling mechanism to adjust the diffusion time step in segments. This mechanism enhances semantic constraints in the early high-noise stage and emphasizes texture and edge modeling in the later low-noise stage, thereby improving visual quality while ensuring semantic consistency. The final generated image is restored to RGB space by the VAE decoder and uniformly output in 512×512 resolution PNG format.

The system is implemented using PyTorch, and all module structures and reasoning processes are highly modularized, which facilitates subsequent mechanism expansion and function replacement. The overall architecture takes into account both semantic expression accuracy and image generation quality while maintaining flexibility, providing a unified generation framework support for multi-style semantic control tasks.

SEMANTIC EMBEDDING AND CONDITIONAL GUIDANCE MECHANISM

The system uses the pre-trained CLIP model as a text semantic encoder and uses its text encoding path to extract the semantic features of the input text. The text is input into the CLIP text encoder in natural language form, enters the embedding layer after word segmentation, and then is processed by the Transformer backbone network to output a fixed-length global semantic feature vector c:

$$c = \text{TextEncoder}(T), \quad c \in \mathbb{R}^d \tag{1}$$

d is the semantic embedding dimension. This vector does not undergo any further semantic aggregation operations and is directly used as a global conditional semantic vector to input the conditional control branch of the diffusion model to guide the image generation process.

After the semantic features are output, they are connected to a multi-layer fully connected linear mapping layer for dimensional adjustment to match the key and value vector dimension requirements of the cross attention module in the UNet architecture. The linear mapping weight is W_c and the bias is b_c, then the converted semantic vector is:

$$c' = W_c \cdot c + b_c \tag{2}$$

The adjusted vector is replicated and expanded in the sequence dimension, and the length corresponds to the number of tokens in the spatial feature map, forming the conditional semantic matrices K, V. This matrix is input as the key and value into the cross-attention module of each layer of UNet, and participates in the attention calculation with the query vector Q:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

The UNet structure is divided into a downsampling path, an intermediate bottleneck path, and an upsampling path. Conditional semantics are injected at each stage to achieve multi-scale semantic guidance. The cross-attention module is connected in parallel with the basic convolutional residual module to form a dual-branch structure. The cross-attention module receives the current layer image features as the query vector and the semantic embedding as the key and value vectors, and obtains the modulation features after

calculating the attention weights. The modulated features are residually connected with the original image features, and the output enters the next processing layer. Figure 1 shows the response distribution of the cross-attention mechanism in different areas of the image under the guidance of semantic conditions. It can be seen that the model is formed in the sky and beach areas of the image, reflecting the response characteristics to the semantics of the input text.

The semantic conditional insertion modular package includes a semantic buffer, a conditional dimension transformer, and a multi-layer injection controller. The semantic buffer fixes the semantic vector after the input text is encoded and does not participate in subsequent dynamic updates. The dimension transformer linearly projects the semantic embedding and replicates it to a fixed sequence length. The injection controller manages the conditional enablement status of each cross-attention module, enabling or disabling semantic injection at different levels.

The image feature query vector is composed of the spatial features of the corresponding layers of each cross-attention module in UNet. The cross-attention module is executed before the convolution residual path at each layer, and the calculated semantic attention results are incorporated into the image features in a residual manner. During the image generation process, conditional control realizes semantic adjustment at each time step and resolution layer. The number of semantic vector copies matches the number of feature map tags, and the width of the conditional pathway in the cross-attention module is consistent with the image branch to strengthen the spatial semantic consistency constraint.

The temporal context information in the diffusion sampling process is injected into UNet synchronously as a condition. The temporal information is encoded by the sine and cosine position $\phi(t)$ and sent to the multi-layer linear mapping and activation function to form a fixed-length time vector τ:

$$\tau = \sigma(W_t \cdot \phi(t) + b_t) \tag{4}$$

The time vector and semantic embedding are concatenated into a joint conditional vector after dimensionality unification:

$$z_{\text{cond}} = [c' || \tau] \tag{5}$$

This vector is processed as a conditional input using a unified mapping in each layer of the cross-attention module to adjust the expression strength of semantic control at different sampling stages.

OPTIMIZATION DESIGN OF CFG STRATEGY

In the implementation of the CFG strategy, this paper adopts the linear interpolation method of the inference results of the conditional and unconditional diffusion models, aiming to improve the semantic consistency and diversity of the generated images. During the sampling process, the model simultaneously executes two forward propagation paths with and without conditional input, obtaining conditional prediction noise $\hat{\epsilon}_{\text{cond}}$ and unconditional prediction noise $\hat{\epsilon}_{\text{uncond}}$ respectively. The two are weightedly fused to produce the final prediction noise:

$$\hat{\epsilon}_t = \omega_t^{(g)} \cdot \hat{\epsilon}_{\text{cond}} + \left(1 - \omega_t^{(g)}\right) \cdot \hat{\epsilon}_{\text{uncond}} \tag{6}$$

$\omega_t^{(g)}$ is the semantic guidance weight coefficient of step t. This weight coefficient is based on a fixed value and introduces a dynamic adjustment mechanism at a specific sampling stage to enhance the model's adaptability to semantic information at different sampling time steps. Figure 2 shows the dynamic change trend of the guidance weight during the sampling process, as well as the gradual clarity of the image corresponding to the key time step.

The dynamic adjustment mechanism is designed as a function mapping between sampling time steps and weighting coefficients. For this purpose, $\omega_t^{(g)}$ is defined as a piecewise linear function:

$$\omega_t^{(g)} = \begin{cases} \omega_{\max}, & t \le t_0, \\ \omega_{\max} - \lambda(t - t_0), & t_0 < t < t_1, \\ \omega_{\min}, & t \ge t_1, \end{cases} \tag{7}$$

t_0 and t_1 are the time steps for the start and end of linear attenuation, $\omega_{\max}$ and $\omega_{\min}$ represent the upper and lower limits of the guidance strength, and λ is the attenuation rate.

The weighting coefficient is kept at a high level at the beginning of sampling to enhance the guiding effect of semantic information on the rough structure; as the sampling progresses, the weighting strength is gradually reduced to reduce the constraints of semantic information on detail generation. The mapping function uses a piecewise linear attenuation strategy, and the parameters are adjusted comprehensively according to the semantic matching degree and image quality indicators of the samples generated during training, and finally solidified for use in the inference stage.

The dynamic adjustment of the guided weights is combined with a multi-scale scheduling mechanism of the noise scale. This mechanism provides the time step adjustment σ_t of the noise variance during the sampling process, which in turn affects the adjustment function of $\omega_t^{(g)}$:

$$\omega_t^{(g)} = f(\sigma_t) = \omega_{\max} \cdot \exp\left(-\beta \cdot \sigma_t^2\right) \tag{8}$$

β is a hyperparameter that controls the exponential decay rate. This function enhances the coupling relationship between semantic guidance weight and noise level, and refines the semantic control granularity of each stage of image generation. In order to reduce the instability of training and reasoning caused by noise weighting in CFG, a regularization mechanism of conditional and unconditional noise prediction results is introduced to measure the relative difference between the outputs of the two paths. This mechanism calculates the Euclidean distance $\mathcal{L}_{\text{dist}}$ and cosine similarity $\mathcal{L}_{\cos}$:

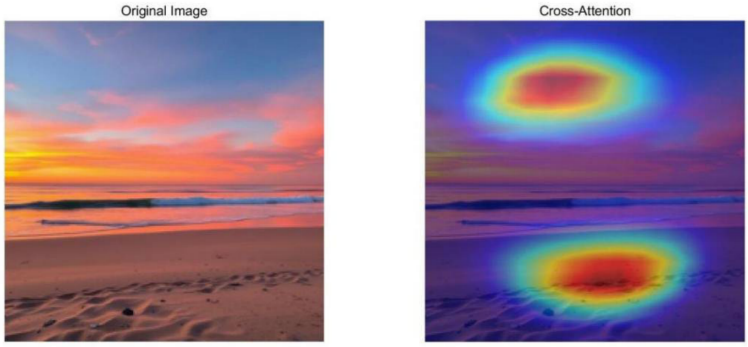

FIGURE 1. Image attention response guided by conditional semantics

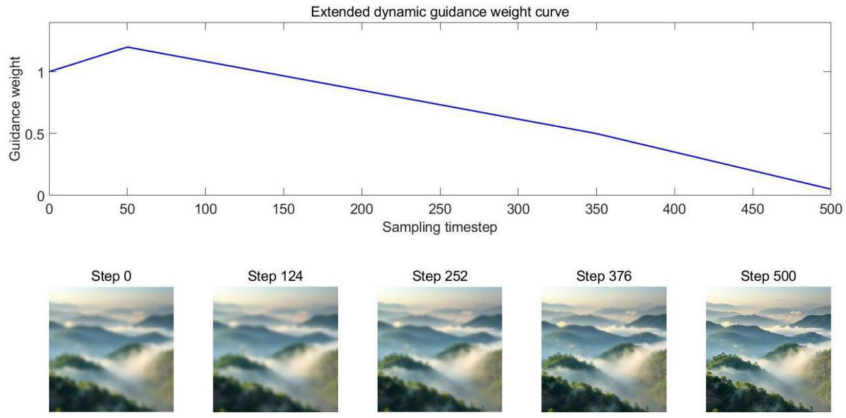

FIGURE 2. Schematic diagram of image denoising evolution under dynamic CFG scheduling

$$\mathcal{L}_{\text{dist}} = \|\hat{\epsilon}_{\text{cond}} - \hat{\epsilon}_{\text{uncond}}\|_2^2 \tag{9}$$

$$\mathcal{L}_{\cos} = 1 - \frac{\langle \hat{\epsilon}_{\text{cond}}, \hat{\epsilon}_{\text{uncond}} \rangle}{\|\hat{\epsilon}_{\text{cond}}\|_2 \cdot \|\hat{\epsilon}_{\text{uncond}}\|_2} \tag{10}$$

If the distance or similarity exceeds the preset threshold, the weight adjustment is triggered. The weight adjustment is guided by the sliding average algorithm with momentum:

$$\omega_t^{(g)} \leftarrow \mu \cdot \omega_{t-1}^{(g)} + (1-\mu) \cdot \widetilde{\omega}_t \tag{11}$$

$\mu \in [0,1]$ is the momentum factor, and $\widetilde{\omega}_t$ is the new guide value obtained by the current calculation. To achieve efficient calculation of CFG, the system integrates and optimizes the network forward propagation process of conditional and unconditional prediction noise. The processing branch of conditional input in the UNet network is implemented using conditional batch normalization and conditional residual connection, allowing conditional information to be dynamically embedded while shielding the input of conditional information in the unconditional path. Multi-threaded asynchronous scheduling is introduced to parallelize the calculation of two paths, further shortening the sampling time.

IMPLEMENTATION OF MULTI-SCALE NOISE SCHEDULING MECHANISM

The multi-scale noise scheduling mechanism is based on the multi-level feature scale of the UNet structure and introduces a hierarchical noise injection scheme in the diffusion sampling process. This mechanism uses the scheduling function $\mathcal{S}^{(l)}$ to generate the standard deviation coefficient to control the noise intensity in the feature maps of different spatial resolutions:

$$\sigma_t^{(l)} = \mathcal{S}^{(l)}(t), \quad l \in \{1, 2, \ldots, L\} \tag{12}$$

$\sigma_t^{(l)}$ represents the standard deviation of the noise at the l-th scale in the time step t, and L is the total number of scales used in UNet.

The noise scheduling function is constructed as a set of time step mapping curves, each of which corresponds to a resolution level in the UNet structure, and is defined on a unified time step scale through multi-segment linear interpolation to ensure the smoothness of noise changes with sampling time:

$$\mathcal{S}^{(l)}(t) = \text{interp}\Big(\{(t_i, \sigma_i^{(l)})\}_{i=1}^{N}\Big) \tag{13}$$

t_i is the sampling time step node, and $\sigma_i^{(l)}$ is the corresponding standard deviation value set. Before each sampling calculation, the variance value of each scale noise

is extracted from the scheduling function according to the current time step to generate a Gaussian noise tensor of the corresponding size:

$$\epsilon_t^{(l)} \sim \mathcal{N}\left(0, \left(\sigma_t^{(l)}\right)^2 \cdot I\right) \tag{14}$$

The tensor dimension is consistent with the UNet feature map $F_t^{(l)}$. The noise tensor is superimposed on the corresponding feature map in the encoder and decoder paths of UNet, and is input into the subsequent network layer together with the feature map convolution result:

$$\widetilde{F}_t^{(l)} = F_t^{(l)} + \epsilon_t^{(l)} \tag{15}$$

The encoding path uses high variance noise at each scale to enhance the degree of freedom of structural expression, and the decoding path variance is reduced layer by layer to restore texture details and edge structures. The diffusion time step encoding uses sine and cosine periodic functions:

$$\tau_t = \left[\sin\left(\frac{2\pi t}{T_f}\right), \cos\left(\frac{2\pi t}{T_f}\right)\right] \tag{16}$$

The temporal encoding vector τ_t is provided as input to the scheduling function to generate scale mapping parameters.

The scheduling mechanism introduces a noise amplitude adaptive adjustment module, which combines the conditional embedding semantic strength and the time step position to fine-tune the noise amplitude based on the scheduling function output:

$$\sigma_t^{(l)} \leftarrow \sigma_t^{(l)} \cdot \left(1 + \alpha^{(l)} \cdot \gamma\right) \tag{17}$$

$\sigma_t^{(l)}$ is the semantic guidance amplification factor at each scale, estimated by the conditional encoding strength extracted by CLIP:

$$\gamma = \frac{\|c_{\text{clip}}\|_2}{\|c_{\text{clip}}\|_2 + \|z_t\|_2} \tag{18}$$

c_{clip} is the text conditional encoding vector, and z_t is the current denoising state tensor. A linear mapping relationship is established between the noise fine-tuning coefficient and the guidance intensity, and the mapping function is a threshold function controlled by two parameters.

This mechanism has a parameter coupling relationship with the CFG process. The sampling time step and the scheduling curve position act together on the adjustment logic of the CFG strength. The joint action mechanism is controlled by a joint mapping table $\mathcal{M}$:

$$\omega_{t,l}^{(g)} = \mathcal{M}\left(t, \sigma_t^{(l)}\right) \tag{19}$$

The mapping table selects the corresponding semantic guidance strength coefficient according to the time step position and noise scale, and passes it to the corresponding attention module in UNet to adjust the fusion ratio of the conditional vector and the feature map:

$$F_{\text{attn}}^{(l)} = \omega_{t,l}^{(g)} \cdot F_{\text{cond}}^{(l)} + \left(1 - \omega_{t,l}^{(g)}\right) \cdot F_{\text{uncond}}^{(l)} \tag{20}$$

The feature map comparison module is introduced in the sampling output process to record the distribution of intermediate feature maps at the main scale of each time step and calculate the gradient consistency $\mathcal{G}_t^{(l)}$ and edge density $\mathcal{E}_t^{(l)}$:

$$\mathcal{G}_t^{(l)} = \left\| \nabla F_t^{(l)} - \nabla F_{t-1}^{(l)} \right\|_1 \tag{21}$$

$$\mathcal{E}_t^{(l)} = \max\left(\left|\nabla^2 F_t^{(l)}\right|\right) \tag{22}$$

The comparison module runs synchronously in the generation process and does not affect the main reasoning process. The noise scheduling mechanism is linked to the post-processing module. After the generation is completed, the final noise amplitude output by the scheduling function is used to drive the sharpening and filtering parameter adjustment in the post-processing stage, matching the texture density of the image detail layer and enhancing the local perceptual quality of the output results.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

EXPERIMENTAL DATA AND ART STYLE LABEL SETTING

The training data for this study is based on the public multimodal dataset LAION-SG, which focuses on high-quality text-image pairs. The length of the text description is controlled between 20-40 tokens, and the Byte-Pair Encoding (BPE) word segmentation standard is used. To ensure the semantic guidance effect, the original data is manually screened to remove low-quality samples such as those with vague visual content or one-sided text descriptions. To adapt to the task of generating artistic inspiration, this paper collected an additional 30,000 images covering a variety of artistic styles, including 10 types of labels such as impressionism, surrealism, cyberpunk, ukiyo-e, minimalism, vaporwave, futurism, ink style, pixel art, and graffiti art.

MODEL TRAINING PARAMETERS AND HARDWARE CONFIGURATION

The basic image generation module uses a text-guided diffusion generation network built on a latent space diffusion architecture. The text feature extraction part uses the CLIP model (the structure uses ViT-B/16). The extracted semantic vector is linearly transformed and dimensionally adjusted before being embedded into the conditional diffusion model. The injection position is the cross-layer attention module in the middle layer of the encoder-decoder path. The conditional control method adopts the CFG strategy. In the guidance stage, the semantic input is randomly discarded with a probability of 15% to construct an unconditional

reference path. The guidance factor in the reasoning stage is set to 1.2 to enhance the semantic relevance of the generated results.

The diffusion process adopts the standard diffusion modeling mechanism of forward denoising and backward denoising. The initial noise scheduling strategy is linear β scheduling, and the noise parameter range is set to 1×10-4 to 2×10-2. The total number of steps in the image sampling process is set to 50. No parameter update is performed during the sampling phase. Only the frozen generated network structure is called. The diffusion inversion method used is DDPM (Denoising Diffusion Probabilistic Models) reverse sampling.

The AdamW optimizer is used in the model training process. The initial learning rate is set to 2×10-5, the momentum coefficients $\beta1$ and $\beta2$ are 0.9 and 0.98 respectively, and the weight decay coefficient is 0.01. The training batch size is 16 images per step, gradient accumulation is enabled, and the number of cumulative steps is 4, so that the equivalent total batch size is 64. The training cycle is 100 rounds, the total number of training steps does not exceed 100,000 steps, and the loss function is the weighted sum of the pixel reconstruction error and the semantic guidance error. The loss is calculated in both the original space and the latent space.

The model training environment is a single graphics processing unit NVIDIA RTX 3090 with a video memory capacity of 24GB. The training process uses mixed precision computing to reduce video memory usage, and the automatic mixed precision computing framework is enabled to accelerate the forward and backward propagation process. Multi-card distributed training, model parallelism, or parameter sharding technology is not enabled. Under limited video memory conditions, the activation checkpoint mechanism is enabled to release non-critical nodes in the intermediate calculation graph to relieve video memory pressure.

EXPERIMENTAL SETTING COMPARISON MODEL

This experiment compares the proposed method with three mainstream models, including T-Gate, Promptguided Attention Interpolation via Diffusion (PAID), and Self-Cross Diffusion Guidance, in a unified environment. All methods use the same text condition input and are tested on a unified dataset to ensure the comparability and fairness of the results. The experiment covers multiple subtasks, from semantic responsiveness, style control diversity, the balance between expression consistency and image quality, to the impact of multi-scale noise mechanism on detail enhancement, and finally combines user subjective evaluation and key module ablation to comprehensively evaluate the performance of each method in the inspiration generation process. Through this systematic experimental design, the specific contribution and role of each module and mechanism to the generation effect are deeply explored.

V. PERFORMANCE OF SEMANTIC CONTROL AND IMAGE QUALITY IN INSPIRATION GENERATION

EVALUATION OF SEMANTIC RESPONSIVENESS UNDER CONDITIONAL TEXT GUIDANCE

In order to test the ability of the generative model to accurately respond to text semantics under multi-level semantic control conditions, the experiment constructed five sets of semantic prompt texts in three dimensions: style, spatial structure, and semantic fusion, and generated corresponding image samples based on these prompts. The results of some text generation are shown in Figure 3.

By extracting the semantic embedding between each pair of images and their semantic texts and calculating their similarity in the CLIP space, a quantitative score of semantic consistency is obtained. This setting is designed to simulate the adaptive behavior of the model under different semantic control scenarios and examine its semantic modeling stability and expression ability under a unified evaluation index. The comparison of the results under three types of semantic conditions is shown in Figure 4.

The results show that under style control, the scores of this method in all five prompt groups remain between 0.76 and 0.81, which is higher than other methods overall. In the spatial structure task, the score of this method is also in the high range, with an average of 0.78. In the semantic fusion task, this method obtains the highest score under all prompts, with the highest score reaching 0.83.

In order to more comprehensively verify the matching degree between the generated text and the input prompt from the two levels of language consistency and semantic accuracy, a multi-dimensional evaluation system covering three mainstream text evaluation indicators, BLEU (Bilingual Evaluation Understudy), BERT (Bidirectional Encoder Representation from Transformers) scores and CIDEr (Consensus-based Image Description Evaluation), was constructed. This experiment measures the model's vocabulary overlap, deep semantic similarity, and overall content expression ability in the generated descriptions to observe the model's comprehensive performance in various dimensions. The results are shown in Figure 5.

The proposed method achieved the highest score of 0.92 on BERTScore, showing strong semantic consistency, and also reached 0.76 and 0.81 on BLEU and CIDEr. In comparison, T-Gate scored 0.72 on CIDEr, but BERTScore and BLEU were 0.76 and 0.68 respectively; PAID scored 0.72 on BLEU, and the other two were 0.83 and 0.75 respectively. Self-Cross' scores were relatively close, and the overall distribution was relatively even, but the highest score was only 0.77, and it did not have an advantage in any single item.

DIVERSITY AND CONSISTENCY OF STYLE SEMANTIC CONTROL

In order to evaluate the semantic consistency and stability of the model's generation results under different image style control scenarios, this experiment designed a variety

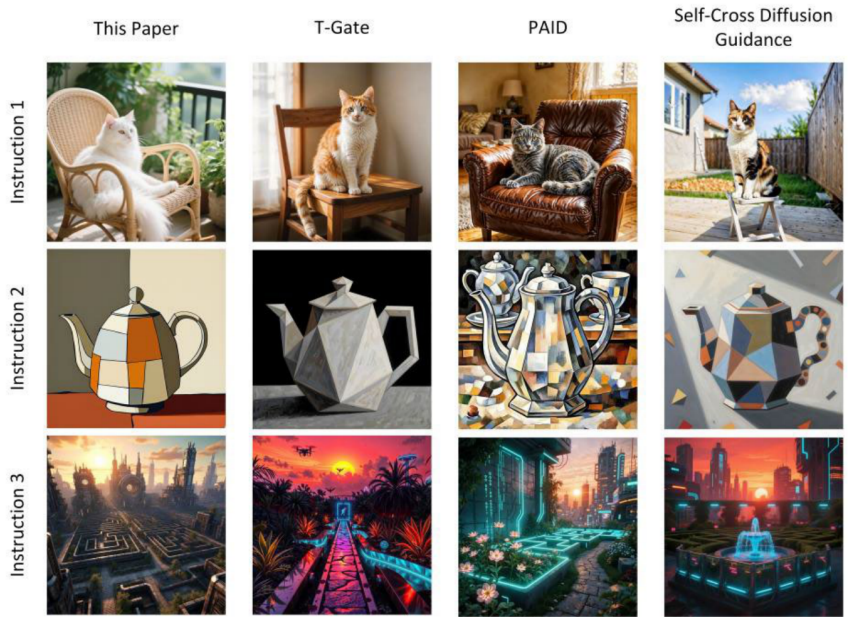

FIGURE 3. Comparison of the generation results of each method under different text inputs

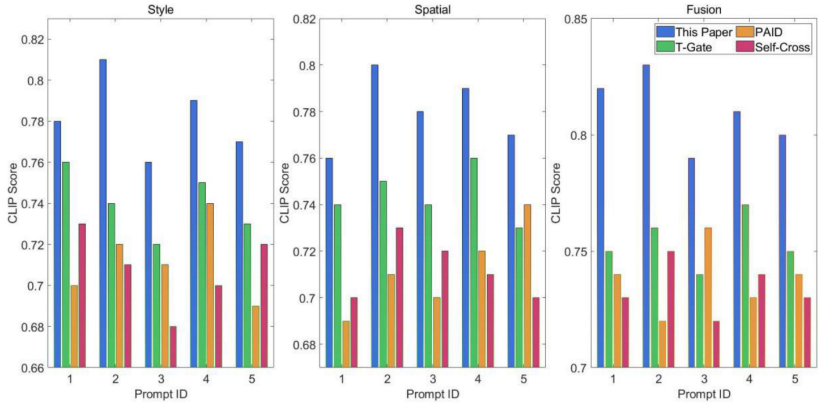

FIGURE 4. Comparison of semantic consistency performance of prompt groups based on CLIP scores under three types of semantic conditions

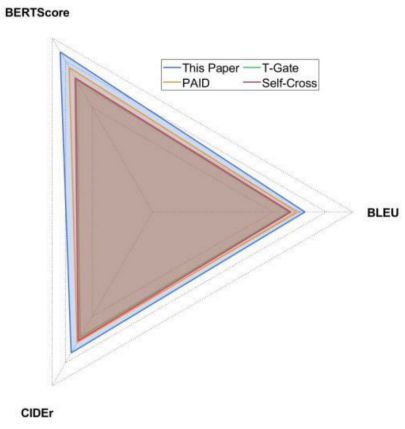

FIGURE 5. Multi-dimensional evaluation of text generation ability based on BLEU, BERTScore and CIDEr indicators

of image-text generation tasks under three types of style conditions: cartoon, oil painting, and sketch, and used the CLIP score as the core indicator to measure the degree of semantic matching between the generated image and the text. This paper sets 10 groups of samples for each style and measures the CLIP distribution of the output results of each

method under the same conditions to test its coordination ability between style preservation and semantic expression. The relevant distribution is shown in Figure 6, which shows the numerical distribution and stability of the CLIP score dimension of each method under three types of styles.

In all styles, the proposed method shows a higher median and a smaller distribution range in CLIP scores. The scores of cartoon style are concentrated between 0.78-0.82, the oil painting style is slightly lower than the cartoon, and the sketch style has a lower overall score and a more dispersed distribution. The box of the Self- Cross method is longer, indicating that it is not stable enough.

BALANCE BETWEEN EXPRESSION CONSISTENCY AND IMAGE QUALITY

In order to further verify the synergistic performance between semantic consistency and image quality of the generated results, the experiment statistically analyzed the results of each method in terms of expression consistency and aesthetic quality, and presented their distribution trends in the form of a two-dimensional scatter plot, so as to observe the degree to which the model takes into account image details and visual perception while maintaining semantic accuracy. The results are shown in Figure 7.

The expression alignment scores of the proposed method are concentrated around 0.81, and the aesthetic quality scores are about 0.79 on average. The overall distribution is relatively concentrated and positioned at the top, showing a good balance between semantic consistency and image quality. In contrast, the scores of other methods are slightly lower, showing a compromise difference in the two indicators. The main reason behind this performance difference is the collaborative optimization mechanism of semantic understanding and image generation quality in model design. The proposed method adopts a more precise alignment strategy when encoding expression semantics, ensuring the accurate transmission of semantic content in the generated image. At the same time, the introduced detail enhancement module effectively improves the aesthetic expression of the image, so that high scores can be achieved in both expression alignment and visual quality. Other methods may have certain deficiencies in semantic capture or lack of balance in image detail expression, resulting in overall low scores.

ROLE OF MULTI-SCALE NOISE MECHANISM IN IMAGE DETAIL ENHANCEMENT

In order to further explore the impact of multi-scale noise mechanism on image detail expression, this experiment designed a series of samples to quantitatively evaluate the three key indicators of edge strength, texture complexity and contrast, aiming to analyze the changes in these detail features before and after the introduction of multi-scale noise. The edge strength uses the Sobel operator to detect the edge of the image and calculate its average response. The texture complexity calculates the contrast feature based on the gray level co-occurrence matrix. The contrast uses the standard deviation of the image gray value. All indicators are normalized and the result ranges from 0 to 1, which is convenient for comparative analysis between different indicators and methods. The results are shown in Figure 8.

In each sub-figure, the blue line represents the result of introducing the multi-scale noise mechanism, and the orange line represents the baseline performance without the introduction of the mechanism. From the data trend, the blue curve is higher than the orange curve in all three indicators. In the edge strength item, the scores of samples with the mechanism are stably distributed between 0.72 and 0.8, while those without the mechanism fluctuate between 0.62 and 0.69; in terms of texture complexity, the scores of samples with the noise mechanism are above 0.7; the scores of samples with noise contrast are between 0.75 and 0.84, while those without the noise mechanism are only between 0.66 and 0.75.

VERIFICATION OF CREATIVE INSPIRATION EFFECT UNDER USER SUBJECTIVE EVALUATION

In order to explore the subjective effectiveness of generated images in the process of creative expression and conception, the experiment invited multiple participants to evaluate a set of different image generation results, and scored them on multiple dimensions such as inspiring creativity, expression fit, and overall perceived satisfaction. The evaluation dimensions cover whether it helps to conceive creative ideas, whether it is inspiring, consistency with the expression intention, the novelty and idea triggering brought by the image, and overall satisfaction. Each score is collected through a five-point Likert scale, and the average scores and standard deviations of different methods are statistically summarized to quantify the inspirational potential and user recognition of each method at the subjective level. The relevant data are summarized in Table 1.

Table 1 evaluates the effect of generated images on supporting users' creative thinking and expression intentions from multiple perspectives. In the dimension of creative conception, the mean of this method is 4.32 and the standard deviation is 0.41, which is higher than 4.16 of T-Gate, 4.03 of PAID and 3.98 of Self-Cross. In the dimension of inspiration, the score of this method is 4.33, slightly higher than the other three. In terms of consistency of expression intention, the proposed method scored 4.35; in the dimensions of image novelty and idea triggering, the proposed method scored 4.34, which performed well. In terms of overall satisfaction, the proposed method scored 4.49 with a standard deviation of 0.38, while T-Gate, PAID, and Self-Cross scored 4.11, 3.89, and 3.74, respectively, with slightly higher standard deviations, reflecting the consistency differences in user feedback.

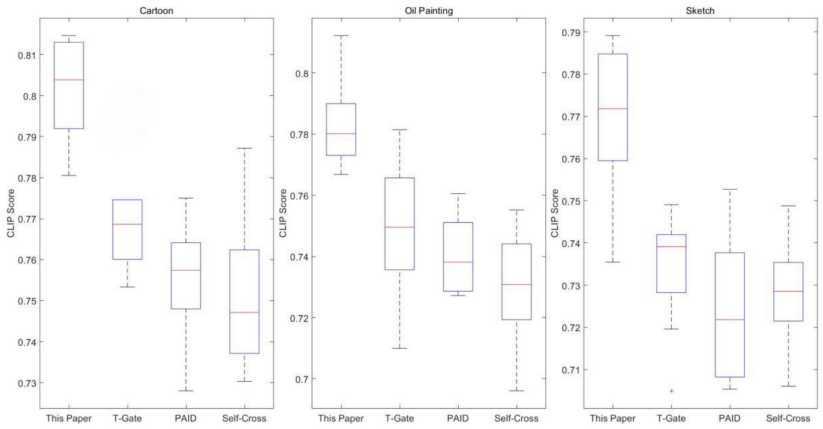

FIGURE 6. Comparison of CLIP consistency distribution under different image style control

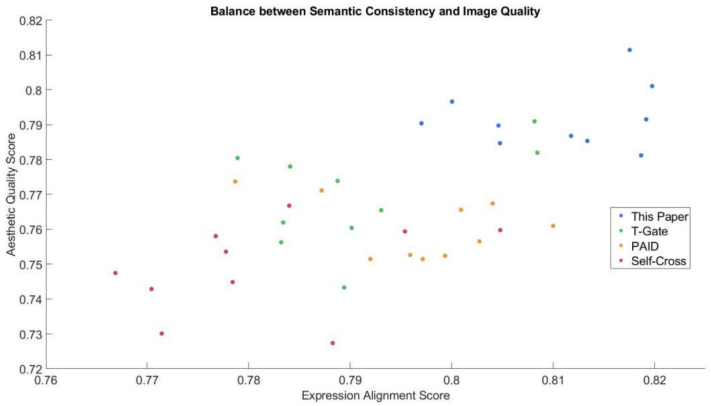

FIGURE 7. Distribution of collaborative performance of expression consistency and image quality

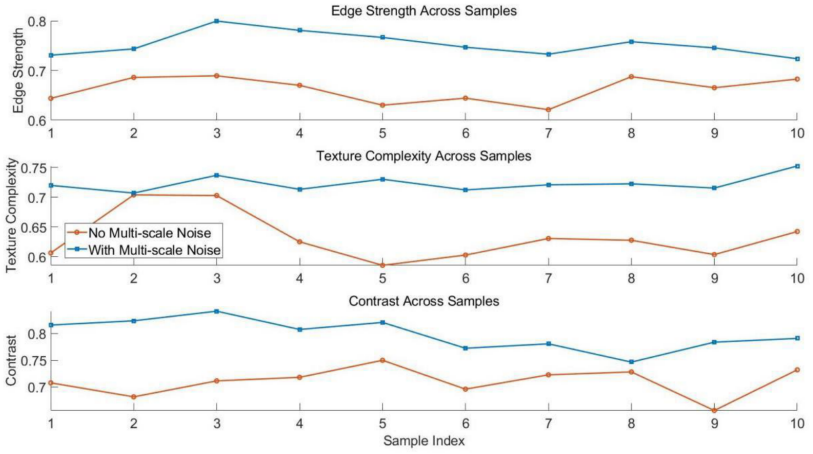

FIGURE 8. Analysis of the impact of multi-scale noise mechanism on image detail features

ABLATION COMPARISON OF KEY COMPONENTS OF THE METHOD

In order to understand the specific impact of each key component of the model on the overall performance, this experiment systematically removed and replaced the multi-scale noise mechanism, CLIP semantic embedding module and Classifier-Free Guidance, aiming to quantify the independent contribution of each module in maintaining semantic consistency, style control ability and image quality. By removing or replacing these core components, their roles and mutual influences in the model generation process are revealed, thus providing data support and theoretical basis for optimizing the model architecture. Table 2 shows the performance of evaluation indicators in three dimensions:

TABLE 1. Statistics of mean and standard deviation of scores of various methods in the creative inspiration dimension under user subjective evaluation

Evaluation Dimension	Method	Mean	Standard Deviation
Helpfulness for Creative Thinking	This paper	4.32	0.41
	T-Gate	4.16	0.48
	PAID	4.03	0.54
	Self-Cross	3.98	0.59
Inspirational Value	This paper	4.33	0.40
	T-Gate	4.05	0.49
	PAID	4.01	0.47
	Self-Cross	4.02	0.63
Alignment with Intended Expression	This paper	4.35	0.42
	PAID	3.96	0.51
	T-Gate	3.92	0.50
	Self-Cross	3.97	0.55
Novelty and Idea Activation from Images	This paper	4.34	0.45
	T-Gate	4.03	0.44
	PAID	3.86	0.50
	Self-Cross	3.99	0.62
Overall Satisfaction	This paper	4.49	0.38
	T-Gate	4.11	0.46
	PAID	3.89	0.52
	Self-Cross	3.74	0.58

semantic consistency, style control, and image quality under each configuration.

The "style control score" in this paper is defined as the semantic consistency measure between the generated image and its target artistic style category. In the experiment, standardized style description text templates (such as "An Impressionist Painting") were constructed for 10 artistic styles (such as Impressionism, Cyberpunk, Ink Painting, etc.), and the pre-trained CLIP-ViT-B/16 model was used to extract the text embedding vectors of each style. For each generated image, the cosine similarity between its CLIP image embedding and 10 styles of text embedding is calculated, and the maximum value is taken as the style matching score of the image. The final style control score is the average of the scores of all 500 generated images in the test set. This method avoids relying on pixel-level features or classifier fine-tuning, and directly evaluates the degree of style alignment in cross-modal semantic space, which is consistent with the generation paradigm with semantic guidance as the core in this paper.

All FID scores reported in this paper are calculated based on the self-built test set of artistic inspiration generation. The test set contains 500 untrained text-image pairs, and the image content covers 10 artistic styles (such as Impressionism, Cyberpunk, Ink and Wash, etc.), and is manually screened to ensure visual quality and semantic clarity. In order to meet the requirements of FID calculation, some images extracted from the test set constitute the real image distribution, and the images generated by each model under the corresponding 500 text prompts are used as the composite distribution, which is input into the pre-trained Inception-v3 network (trained on ImageNet without updating the weights) to extract features, and then the FID value is calculated. This setting ensures that the FID evaluation is highly consistent with the task scene (artistic image generation) in this paper, and avoids the possible deviation caused by the evaluation on the general natural image data set (such as COCO).

The complete model achieved a CLIP similarity score of 0.82 for semantic consistency, a style control score of 0.76, and an FID value of 18.5 for image quality, which is a relatively balanced performance. After removing the multi-scale noise mechanism, the semantic consistency dropped to 0.78, the style control dropped to 0.70, and the FID increased to 23.1, indicating that the image details and overall quality have been weakened. After removing the CLIP semantic embedding, the semantic consistency drops more significantly, to 0.72, but the style control score remains relatively unchanged at 0.74, with an FID of 20.7. Canceling the Classifier-Free Guidance configuration further drops the style control score to 0.67, and the FID rises to 24.5, reflecting that the module is able to maintain image diversity and style stability. Replacing multi-scale noise with single-scale noise drops the semantic consistency and style control to 0.79 and 0.69, respectively, with an FID of 22.0. Finally, when the CLIP encoder is frozen, the semantic consistency is 0.70, the style control is 0.73, and the FID is 21.3, which is lower than the complete model.

From the mechanism level, the change in semantic consistency is mainly affected by the CLIP semantic embedding module, because this module provides a strong ability to align text and image semantics. Removing or freezing this module can lead to a decrease in the model's ability to understand and express semantic information. The multi-scale noise mechanism plays a key role in enriching image details and enhancing texture expression. Its absence or replacement with single-scale noise leads to a decline in style control and image quality, indicating that this mechanism can effectively improve the layering and realism of generated images. The Classifier-Free Guidance module focuses more on adjusting the balance between diversity and consistency in the generation process. After removing this module, the style control score and FID performance

TABLE 2. Ablation comparison

Module Configuration	Semantic Consistency (CLIP Similarity)	Style Control (Style Consistency Score)	Image Quality (FID)
Full Model	0.82	0.76	18.5
Without Multi-scale Noise	0.78	0.70	23.1
Without CLIP Semantic Embedding	0.72	0.74	20.7
Without Classifier-Free Guidance	0.75	0.67	24.5
Replaced with Single-scale Noise	0.79	0.69	22.0
Replaced with Frozen CLIP Encoder	0.70	0.73	21.3

fluctuate significantly, indicating that it plays an irreplaceable role in maintaining the stability and quality of image style. Overall, these experimental data reflect the important contribution of each module of the model to the final generation effect in collaborative work.

COMPARATIVE ANALYSIS WITH EXISTING ADVANCED TEXT-TO-IMAGE MODELS

In order to comprehensively evaluate the competitiveness of the proposed system in the task of artistic inspiration generation, this paper directly compares it with several representative advanced text-to-image generation models, including closed-source business model and mainstream open-source diffusion model.

Specific comparison models include: DALL·E 2, Stable Diffusion v1.5, Stable Diffusion XL1.0 and Imagen. All comparison models use the same 500 test tips (covering abstract concepts, style mixing and scene description) to generate images on the basis of their official open interfaces or open source weights, and calculate indicators under a unified evaluation protocol.

The evaluation index includes three core dimensions: (1) Semantic consistency: using CLIP-ViT-B/32 to calculate the cosine similarity between the generated image and the input text; (2) Style control ability: follow the style control score defined in this paper (maximum value of CLIP matching based on 10 artistic style templates); (3) Image quality: Calculate FID(Inception-v3 feature) on the self-built art test set. All generated images are uniformly adjusted to the resolution of 512×512, and the sampling steps are set according to the recommendations of each model (DALL·E 2 and Imagen use the default API parameters, and SD v1.5 and SDXL use 50-step DDIM sampling).

The results of quantitative comparison with the advanced text-to-image model are shown in Table 3.

Table 3 shows the performance of each model on three indicators. The results show that although DALL·E 2 and Imagen perform well in the general image generation task, they are significantly weaker than this method in the specific dimension of artistic style control (the score of style control is 0.61 and 0.63, respectively), reflecting that their training data are biased towards natural images and their generalization ability to minority artistic styles is limited. Although Stable Diffusion v1.5 supports style hints, it lacks fine-grained semantic alignment mechanism, and the style score is only 0.68. SDXL has improved the image quality and semantic consistency (FID=20.1, CLIP=0.80), but the style control (0.71) is still lower than that of this system (0.76). While maintaining high semantic consistency (0.82) and good image quality (FID=18.5), this method shows obvious advantages in precise control of artistic style, which verifies the effectiveness of the designed multi-scale noise scheduling and CLIP conditional guidance mechanism in artistic creation scenes. Further observation shows that DALL·E 2 and Imagen tend to generate realistic or photographic images, and even if the prompt clearly requires "ink painting style" or "ukiyo-e painting", their output still retains a strong sense of reality and the style characteristics are weakened; Although SD series models can respond to style keywords, they are prone to semantic conflicts or details confusion under complex mixed styles (such as "Cyberpunk+Ukiyo-e"). The system in this paper can more accurately integrate style elements and semantic content, and generate images with artistic expression and conceptual consistency, which is more in line with the creator's demand for visual metaphor and style experiment in the conception stage.

TABLE 3. Quantitative comparison results with advanced text-to-image models

Model	Semantic Consistency (CLIP Similarity)	Style Control (Style Consistency Score)	Image Quality (FID)
DALL·E 2	0.84	0.61	17.8
Imagen	0.85	0.63	16.9
Stable Diffusion v1.5	0.78	0.68	22.3
Stable Diffusion XL 1.0	0.80	0.71	20.1
Ours (Full Model)	0.82	0.76	18.5

VI. CONCLUSION

In this paper, a text-guided diffusion model system for artistic inspiration generation is proposed. Its core technical contribution lies in integrating the semantic embedding depth of CLIP into UNet multi-layer cross-attention module to achieve fine-grained semantic control, introducing a dynamically optimized classifier-free guidance (CFG) strategy to balance the generation diversity and semantic accuracy, and designing a multi-scale noise scheduling mechanism to jointly optimize image details, style expression and generation stability. Experiments show that this method achieves a CLIP semantic similarity of 0.82, a style control score of 0.76 and a FID value of 18.5 on the test set, which is significantly better than the existing advanced models in artistic style control. At the same time, in the subjective evaluation of users, its creative inspiration and overall satisfaction reached 4.33 points and 4.49 points (5 points) respectively, which fully verified the practical value and expression potential of the system in the stage of assisting art conception, and provided an effective path for building an efficient, controllable and artistically expressive human-computer collaborative creative tool.

AUTHOR CONTRIBUTIONS

Conceptualization-Jie Zhou, Yu Zhou and Lu Yang; methodology-Jie Zhou, Yu Zhou and Lu Yang; formal analysis-Jie Zhou, Yu Zhou and Lu Yang; investigation-Jie Zhou, Yu Zhou and Lu Yang; resources-Jie Zhou, Yu Zhou and Lu Yang; writing-Jie Zhou, Yu Zhou and Lu Yang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by project of science and technology research program of Chongqing Education Commission of China, titled"Media Transformation of Light Art and Its Application in Chongqing Public Space"(No.KJQN202101003).

ANIMAL RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Chongqing Normal University. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] V. El Ardeliya, J. Taylor, and J. Wolfson, "Exploration of artificial intelligence in creative fields: Generative art, music, and design," International Journal of Cyber and IT Service Management, vol. 4, no. 1, pp. 40-46, 2024, doi: 10.34306/ijcitsm.v4i1.149.

[2] Y. Li and Q. Zhang, "The Analysis of Aesthetic Preferences for Cultural and Creative Design Trends under Artificial Intelligence," IEEE Access, vol. 12, pp. 158799-158808, 2024, doi: 10.1109/ACCESS.2024.3486031.

[3] M. Monser and E. Fadel, "A modern vision in the applications of artificial intelligence in the field of visual arts," International Journal of Multidisciplinary Studies in Art and Technology, vol. 6, no. 1, pp. 73-104, 2023, doi: 10.21608/ijmsat.2024.274900.1021.

[4] A. Oksanen, A. Cvetkovic, N. Akin, R. Latikka, J. Bergdahl, Y. Chen, et al., "Artificial intelligence in fine arts: A systematic review of empirical research," Computers in Human Behavior: Artificial Humans, vol. 1, no. 2, Art. no. 100004, 2023, doi: 10.1016/j.chbah.2023.100004.

[5] B. Wang, Q. Chen, and Z. Wang, "Diffusion-based visual art creation: A survey and new perspectives," ACM Computing Surveys, vol. 57, no. 10, pp. 1-37, 2025, doi: 10.1145/3728459.

[6] N. Dehouche and K. Dehouche, "What's in a text-to-image prompt? The potential of stable diffusion in visual arts education," Heliyon, vol. 9, no. 6, Art. no. e16757-e16769, 2023, doi: 10.1016/j.heliyon.2023.e16757.

[7] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," ACM Transactions on Graphics (TOG), vol. 41, no. 4, pp. 1-11, 2022, doi: 10.1145/3528223.3530104.

[8] R. Fridman, A. Abecasis, Y. Kasten, and T. Dekel, "Scenescape: Text-driven consistent scene generation," in Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S. editor. Advances in Neural Information Processing Systems. Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), 10-16 December 2023, Red Hook, NY, USA. Curran Associates Inc, 2023, pp. 39897-39914, [Online]. Available: https://dl.acm.org/doi/10.52202/075280-1734.

[9] R. Safta, "Artistic Inspiration and Divine Inspiration," Limba Şi Literatura–Repere Identitare În Context European, vol. (32), pp. 175-186, 2023.

[10] C. Chang, "Being inspired by media content: Psychological processes leading to inspiration," Media Psychology, vol. 26, no. 1, pp. 72-87, 2023, doi: 10.1080/15213269.2022.2097927.

[11] F. K. El-Shafey, E. L. Mohamed, A. M. Fouad, A. S. Elkhayat, and A. G. Hassabo, "The Influence of Nature on Art and Graphic Design: The Connection with Raw Materials and Prints," Journal of Textiles, Coloration and Polymer Science, vol. 21, no. 2, pp. 385-396, 2024, doi: 10.21608/jtcps.2024.259042.1276.

[12] D. Gamberini, ""Unjustly Tormented by Love": Eros as a Source of Artistic Inspiration in an Epigram for Gian Giorgio Lascaris, Alias Pyrgoteles," Source: Notes in the History of Art, vol. 41, no. 3, pp. 176-185, 2022, doi: 10.1086/720923.

[13] F. Tao, "A new harmonisation of art and technology: Philosophic interpretations of artificial intelligence art," Critical Arts, vol. 36, no. 1-2, pp. 110-125, 2022, doi: 10.1080/02560046.2022.2112725.

[14] L. Long, C. Xinyi, N. Ruoyu W E, T. J. J. Li, and L. C. Ray, "Sketchar: supporting character design and illustration prototyping using generative AI," Proceedings of the ACM on Human-Computer Interaction, vol. 8, no. CHI PLAY, pp. 1-28, 2024, doi: 10.1145/3677102.

[15] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," Journal

of Machine Learning Research, vol. 23, no. 47, pp. 1-33, 2022, doi: papers/v23/21-0635.html.

[16] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," IEEE Transactions on Patern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10850-10869, 2023, doi: 10.1109/TPAMI.2023.3261988.

[17] D. H. Kwon, "Analysis of prompt elements and use cases in image-generating AI: Focusing on Midjourney, Stable Diffusion, Firefly, DALL·E," Journal of Digital Contents Society, vol. 25, no. 2, pp. 341-354, 2024, doi: 10.9728/dcs.2024.25.2.341.

[18] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," in Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S. editor. Advances in Neural Information Processing Systems. Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), 10-16 December 2023, Red Hook, NY, USA. Curran Associates Inc, 2023, pp. 66923–66939, [Online]. Available: https://dl.acm.org/doi/10.52202/075280-2923.