PeerJ Computer Science

Design of personalized creation model for cultural and creative products based on evolutionary adaptive network

Dailei Hu[1], Enshi Wang[1] and Muddassira Arshad[2]

[1] School of Digital Art, Wuxi Vocational College of Science and Technology, WuXi, Jiangsu, China
[2] Department of Software Engineering, University of the Punjab, Lahore, Pakistan

ABSTRACT

In the realm of personalized cultural and creative product design, the capacity for nuanced semantic expression and refined style modulation in image content exerts a pivotal influence on user experience and the perceived creative value. Addressing the limitations of current generative models—particularly in maintaining stylistic coherence and accommodating individualized preferences—this article introduces a novel image synthesis framework grounded in a synergistic mechanism that integrates text-driven guidance, adaptive style modulation, and evolutionary optimization: Evolutionary Adaptive Generative Aesthetic Network (EAGAN). Anchored in the Stable Diffusion architecture, the model incorporates a semantic text encoder and a style transfer module that will realize the image style transfer, augmented by the Adaptive Instance Normalization (AdaIN) mechanism, to enable precise manipulation of stylistic attributes. Concurrently, it embeds an evolutionary optimization component that iteratively refines cue phrases, stylistic parameters, and latent noise vectors through a genetic algorithm, thereby enhancing the system's responsiveness to dynamic user tastes. Empirical evaluations on benchmark datasets demonstrate that EAGAN surpasses prevailing approaches across a suite of metrics—including Fréchet inception distance (FID), CLIPScore, and Learned Perceptual Image Patch Similarity (LPIPS)—notably excelling in the harmonious alignment of semantic fidelity and stylistic expression. Ablation studies further underscore the critical contributions of the style control and evolutionary optimization modules to overall performance gains. This work delineates a robust and adaptable technological trajectory with substantial practical promise for the intelligent, personalized generation of cultural and creative content, thus fostering the digital and individualized evolution of the creative industries.

Submitted 17 July 2025
Accepted 19 September 2025
Published 14 October 2025

Corresponding author
Enshi Wang, 2056016@wxsc.edu.cn

Academic editor
Muhammad Asif

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj-cs.3288**

Subjects Adaptive and Self-Organizing Systems, Artificial Intelligence, Data Mining and Machine Learning, Data Science, Social Computing

Keywords Personalized generation, Cultural creativity, Diffusion model, Style transfer, Evolutionary optimization

INTRODUCTION

Amid the backdrop of increasingly diversified cultural consumption, personalized cultural and creative products have emerged as a vital conduit bridging traditional heritage with contemporary aesthetics. These products serve not only as vessels of regional identity, historical memory, and national symbolism, but also as expressions of individual taste and aesthetic sensibility. Conventionally, the design of such artifacts has relied heavily on the

subjective intuition and experiential knowledge of professional designers—a process often marked by extended production cycles and limited customization—rendering it inadequate for addressing the nuanced and varied expressive needs of users across diverse cultural and situational contexts. As consumers place growing emphasis on cultural value and unique experiential resonance, market demand for personalized cultural and creative offerings continues to escalate, thereby catalyzing a paradigm shift in design philosophy from mass-oriented production to individualized customization. The design domain is thus compelled to explore new generative frameworks and methodological innovations to actualize this transition. The evolution of personalized cultural and creative design not only signifies a transformation in the modalities of cultural articulation, but also represents a constructive response to the heightened user agency characteristic of the digital era (*Zhang, 2021*). Today's users increasingly eschew passive consumption in favor of active participation, seeking to assert self-identity and cultural preference through tailored design engagements. This emergent trend not only drives service innovation within the cultural and creative industry value chain but also imposes heightened demands on the intelligence and adaptability of design technologies. Against this backdrop, the imperative to construct a system endowed with autonomous creative capabilities—capable of interpreting linguistic input and generating visually diverse content imbued with cultural coherence—has become central to enhancing the efficiency, accessibility, and personalization of creative production (*Li & Wang, 2022*). Accordingly, research into personalized image generation within cultural and creative domains is not only of significant theoretical import but also holds substantial promise for industrial application.

In recent years, the rapid advancement of artificial intelligence—particularly the widespread deployment of deep learning across vision, language, and generative domains—has profoundly reshaped the pathway toward personalized design. Whereas traditional design methodologies have historically depended on human expertise and manual rendering, contemporary intelligent design systems are now capable of autonomously translating user input into visual representations. Personalization modeling has thus emerged as a pivotal research frontier within artificial intelligence, with its principal objective being the generation of content outputs that align closely with individual aesthetic preferences, behavioral patterns, or contextual nuances (*Hui, 2021*). In this transformative process, deep learning technologies occupy a central role, having achieved remarkable breakthroughs especially in image synthesis and style transfer. On the front of image generation, models such as generative adversarial network (GAN), variational autoencoders (VAE), and diffusion models have seen rapid development and now underpin high-fidelity image creation. Diffusion models, in particular, have become a cornerstone of text-to-image generation due to their exceptional stability and fine-grained detail reproduction. Concurrently, techniques such as style-based GAN architecture (StyleGAN) and Adaptive Instance Normalization (AdaIN) have significantly advanced the controllability of image stylization, enabling the same semantic content to be rendered in a multitude of culturally resonant visual languages. Moreover, the advent of multimodal frameworks such as CLIP (Contrastive Language-Image Pretraining) has facilitated deep semantic alignment between textual and visual modalities, rendering the automatic

translation of linguistic expressions into coherent imagery both feasible and effective. The confluence of these technological trajectories not only provides a robust theoretical and algorithmic foundation for user-centric content generation, but also unlocks a vast horizon of possibilities for the intelligent creation of cultural and creative artifacts (*Li & Wang, 2022*).

In the domain of personalized cultural and creative design, the integration of text-to-image (TIG) generation and style transfer technologies offers a highly efficient and imaginative paradigm. TIG empowers users to synthesize image content aligned with semantic intent solely through natural language descriptions, eliminating the prerequisite for manual illustration skills. Style transfer, on the other hand, enables the transformation of generated imagery into culturally resonant artistic forms, thereby facilitating the visual reinterpretation of content within specific cultural frameworks (*Han, Shi & Shi, 2022*). This generation–transformation mechanism is particularly well-suited to the demands of cultural and creative design, where the confluence of semantic precision and stylistic authenticity is paramount, substantially enhancing both creative autonomy and controllability.

Nevertheless, prevailing methodologies often exhibit limited responsiveness to user-specific preferences, resulting in outputs that lack sufficient personalization. Furthermore, existing models frequently struggle to reconcile the inherent trade-off between visual fidelity and stylistic expressiveness, thus impeding the harmonious alignment of semantic accuracy and aesthetic cohesion. In response, this article introduces the Evolutionary Adaptive Generative Aesthetic Network (EAGAN)—a personalized generation framework that fuses textual semantic modeling, style-adaptive modulation, and an evolutionary optimization strategy. By jointly optimizing text prompts, style vectors, and generative parameters through evolutionary algorithms, EAGAN achieves a dynamic synthesis of semantic intent and stylistic form.

The principal contributions of this work are threefold:

(1) A style-adaptive generative architecture is proposed, integrating a diffusion-based backbone with multi-style AdaIN, enabling the resultant imagery to preserve semantic content while supporting personalized stylistic articulation
(2) An evolutionary optimization module is incorporated, employing iterative genetic search to enhance the model's adaptability across varying user preference inputs
(3) A comprehensive multi-source evaluation framework is established, comprising both quantitative metrics and subjective assessments across open benchmarks focused on semantic fidelity and style control, which collectively substantiate the superior performance of the proposed method across multiple evaluation dimensions

The remainder of this article is structured as follows: 'Related Works' reviews related works on text-based image generation and style transfer. 'Materials and Methods' presents the proposed EAGAN framework in detail. 'Experiment Result and Analysis' outlines the experimental setup and provides a comprehensive analysis of the results. 'Discussion' offers an in-depth discussion, and the concluding remarks are provided in the final section.

RELATED WORKS

Text generation image study

In the field of text-to-image generation, GANs have been widely adopted to facilitate multimodal synthesis by generating visual content conditioned on textual descriptions. However, conventional GANs often suffer from instability during training, which can result in mode collapse and suboptimal image quality. To address these challenges, researchers have proposed hierarchical generation architectures. Notably, the Stacked GAN introduced a two-stage generation process to sequentially refine image resolution and detail, with each stage operating independently (*Tominaga & Seo, 2022*). Building on this concept, *Zhang et al. (2019)* presented StackGAN++, which restructured the generation pipeline into a tree-like, end-to-end framework comprising multiple generators and discriminators for enhanced feature refinement. Expanding upon stacked architectures, attention-based GANs emerged to improve the semantic alignment between text and image. Among these, Attentional GAN by *Xu et al. (2020)* and Rich Feature GAN by *Cheng et al. (2020)* introduced attention mechanisms to selectively emphasize word-level semantics during image synthesis. Further advancements include Deep Fusion GANs proposed by *Tao et al. (2022)*, Semantic GANs and Semantic-Spatial Aware GANs introduced by *Liao et al. (2022)*, and Recurrent Affine Transformation GANs developed by *Ye et al. (2024)*, all of which employed conditional affine transformations to achieve deeper integration of textual and visual modalities. Experimental evaluations demonstrate that even single-level architectures utilizing such transformations can produce photorealistic outputs that rival or surpass those generated by complex multi-stage models. However, the omission of word-level semantic granularity in these models often leads to a deficiency in fine-detailed image features.

To overcome this limitation, *Tan et al. (2023)* introduced a fine-grained semantic evaluation metric known as Semantic Similarity Distance, which inspired further development in contrastive pretraining methods. Based on this principle, CLIP model was proposed (*Radford et al., 2021*), laying the groundwork for the Parallel Deep Fusion GAN (PDF-GAN). PDF-GAN leverages multi-level semantic integration, employing a discriminator capable of simultaneously evaluating global coherence and local detail, thereby enhancing the fidelity and semantic richness of the generated images.

Image style transfer study

CariGANs, proposed by *Cao, Liao & Yuan (2020)* leverage neural networks in conjunction with principal component analysis to learn a mapping from facial keypoints in real photographs to those in caricature counterparts, thereby facilitating deformation based on structural correspondence. However, the exaggeration effect in this approach is rigidly anchored to the input image, constraining the generative diversity of resulting caricatures. Nonetheless, the tight coupling between stylization and deformation modules hinders its adaptability across varied application contexts (*Hou et al., 2021*).

Turja et al. (2022) further explored caricature deformation by predicting a deformation field from paired image data. Although effective in capturing artist-guided exaggeration,

this method is heavily reliant on labor-intensive, high-quality supervision and treats caricature generation as a deterministic one-to-one mapping task, thus failing to produce diverse outputs. In contrast, *Hou et al. (2021)* incorporated latent encoding strategies to enrich the diversity of exaggeration patterns. While this approach offers potential for varied deformations, the absence of robust supervision introduces challenges in controllability and frequently leads to unstable or inconsistent distortions. To address these limitations, styleCariGAN (*Jang et al., 2021*) enhances shape exaggeration by embedding dedicated exaggeration blocks into pre-trained StyleGAN layers (*Bermano et al., 2022*). Despite its innovative architecture, the model predominantly emphasizes texture manipulation and struggles with achieving meaningful structural distortion, such as ironic or satirical deformation. Complementarily, *Men et al. (2022)* devised a segment-based, database-driven matching technique wherein user-assisted segmentation is employed to isolate hairstyles and facial regions. These segmented elements are then matched with stylized samples from a curated database, which are recombined to synthesize the final caricatured output (*Men et al., 2022*).

The aforementioned research underscores the growing maturity of cross-modal analysis and style transfer methodologies grounded in deep learning, particularly within the domain of artistic creation. Consequently, in the context of personalized cultural and creative product development, the application of deep learning and cross-modal generative technologies offers substantial potential to shorten production cycles and fulfill the demands of intelligent, user-centric customization. Building upon traditional text-to-image diffusion models, this study incorporates a style transfer module to accommodate more diverse generative requirements and employs a tailored planning algorithm to expand the range of design aesthetics. These advancements hold significant promise for the future of personalized product creation, marking a pivotal step toward efficient, adaptive, and culturally resonant design innovation.

MATERIALS AND METHODS

Text vector and diffusion-based image generation

The image generation module serves as the central component of the proposed system, tasked with translating text-based semantic vectors and style preference vectors into coherent visual representations. In this work, the module employs a generation framework built upon a diffusion-based architecture—specifically, Stable Diffusion—which reconstructs images through an iterative denoising process, thereby enhancing both visual clarity and semantic alignment, while preserving considerable generative flexibility. Given the integral role of user input in the personalized design of cultural and creative products, Stable Diffusion is adopted as the backbone of the image generation module. This model, grounded in the latent diffusion model, achieves high-fidelity image synthesis with superior computational efficiency (*Du et al., 2023*). Unlike earlier diffusion approaches that operate directly in pixel space, Stable Diffusion performs transformations within a

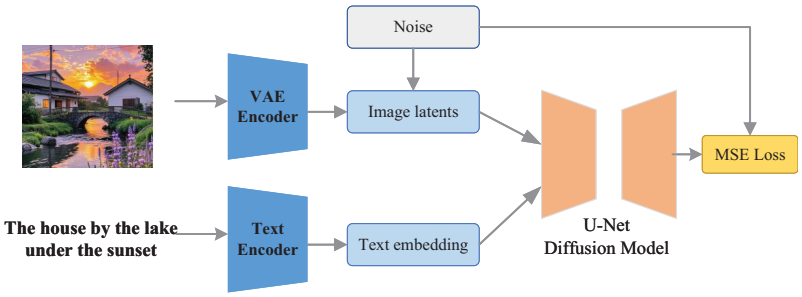

Figure 1 **The framework for the stable diffusion.** Full-size DOI: 10.7717/peerj-cs.3288/fig-1

lower-dimensional latent space, substantially reducing the computational burden associated with training and inference. The overarching architecture is illustrated in Fig. 1.

Stable Diffusion uses a text encoder in a pre-trained CLIP model to extract the semantic representation of the input text. The input text T is tokenized and fed into the Transformer, which outputs the embedding vector $\mathbf{c} \in \mathbb{R}^d$. This semantic vector is fed into the attention module of UNet as conditional information that guides the semantic direction of image generation. Its main direction is represented by Eq. (1):

$$\mathbf{c} = \mathrm{CLIP}_{\mathrm{text}}(T). \tag{1}$$

While the other way of encoding results in the image $x \in \mathbb{R}^{H\times W\times 3}$. It is encoded by VAE as potential representation $z \in \mathbb{R}^{h\times w\times c}$, where $h \ll H, w \gg W$, to reduce the arithmetic burden. The representation result of its main encoder z and decoder x is given in Eqs. (2) and (3):

$$z = \mathrm{Enc}_{\phi}(x) \tag{2}$$

$$\hat{x} = \mathrm{Dec}_{\theta}(z). \tag{3}$$

After the feature extraction of the two-way network is completed, the forward process adds the latent variable z to the Gaussian noise to construct the training samples:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t} \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

Let t denote the discrete time step, and α_t the noise attenuation factor determined by the scheduler. We denote by z_t the noisy latent representation at step t, by ε the Gaussian noise sample to be removed, and by ccc the textual condition embedding provided by the encoder. The conditional UNet network is trained to predict the noise component ε from (z_t, t, c), thereby learning to iteratively denoise the latent representation under text guidance:

$$\hat{\epsilon}_{\theta}(z_t, t, \mathbf{c}) = \mathrm{UNet}_{\theta}(z_t, t, \mathbf{c}). \tag{5}$$

The cross-modal attention mechanism is introduced in UNet for fusing textual semantics, and the time step t is added to the network as an additional condition through

time coding. So far the final loss function is obtained in the form of MSE whose diffusion loss is shown in Eq. (6):

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_0,t,\epsilon}\left[\epsilon - \hat{\epsilon}_\theta(z_t, t, \mathbf{c})^2\right]. \tag{6}$$

Finally, after completing the corresponding training, for a given present description T, the condition vector $\mathbf{c}$ is generated, a latent variable $z_T \sim \mathcal{N}(0, I)$ is sampled from a Gaussian distribution, and the denoising is iterated step by step:

$$z_{t-1} = \text{DenoiseStep}\left(z_t, \hat{\epsilon}_\theta(z_t, t, \mathbf{c})\right). \tag{7}$$

The final iteration is completed to get the reconstructed image as:

$$\hat{x} = \text{Dec}_\theta(z_0). \tag{8}$$

In this study, the image generation module functions as a foundational platform upon which adaptive control mechanisms and evolutionary strategies are integrated, enabling the synthesis of images that are diverse, semantically coherent, and stylistically distinct. As a central architectural element, it plays a critical role in facilitating personalized image generation, effectively aligning output aesthetics with user-defined preferences and contextual demands.

AdaIN-based image style transfer

The style adaptive module is designed to enable controllable and personalized modulation of image styles by incorporating external style representations—such as exemplar images or user-defined preference vectors—to align the visual output with specific cultural aesthetics or individual taste profiles (*Gu & Ye, 2021*). In this work, we employ AdaIN as the core mechanism for style regulation, owing to its efficiency and differentiability in facilitating image style transfer. AdaIN achieves style adaptation by normalizing the feature statistics of the content image—specifically, by adjusting its channel-wise mean and variance—and replacing them with those of the target style image. This operation alters the image's textural and stylistic characteristics while preserving its underlying structural semantics. Formally, for an input image, AdaIN can be expressed as:

$$\text{AdaIN}(f_c, f_s) = \sigma(f_s) \cdot \left(\frac{f_c - \mu(f_c)}{\sigma(f_c)}\right) + \mu(f_s) \tag{9}$$

where $f_c \in \mathbb{R}^{C\times H\times W}$: represents the feature map of the content image on a certain layer; $f_s \in \mathbb{R}^{C\times H\times W}$: represents the feature map of the style image; and for $\mu(\cdot), \sigma(\cdot)$ represents the mean and standard deviation calculation function on the channel dimension. The generation of the style-transferred image is ultimately completed through a decoding process. In this study, the style adaptive module serves as a key mechanism for aligning visual output with users' personalized aesthetic preferences. By enabling explicit style control, it ensures that the generated images not only maintain semantic fidelity to the input text but also embody stylistic characteristics that reflect cultural traditions or

individual tastes. This capability constitutes one of the core components in achieving truly personalized cultural and creative expression.

The establishment for the evolutionary adaptive generative aesthetic network (EAGAN)

To enhance the adaptability of the generated multi-style images to a broader spectrum of product design requirements, this study introduces a genetic algorithm (GA)-based optimization strategy. This approach targets three critical elements within the model pipeline: the textual prompts in the diffusion model training phase, the style content vectors during style transfer, and the latent noise inputs in the final decoding stage. Serving as a pivotal mechanism for adaptive refinement, this module leverages population-based evolution—encompassing fitness evaluation, crossover, and mutation operations—to iteratively improve the quality, diversity, and personalization of the generated outputs.

Construct the initial population $\mathcal{P}_0 = \{P_1, P_2, \ldots, P_N\}$, each individual is a set of generative configurations, where *Pi* is shown in Eq. (10):

$$P_i = \left\{\mathbf{t}_i', \mathbf{s}_i, z_i\right\}. \tag{10}$$

Three of the parameters represent the text embedding vector, style vector and noise variable, and the individual assessment fitness function defined based on the above three elements is shown in Eq. (11):

$$F(P_i) = \alpha \cdot \text{CLIPScore}(x_i, T_i) + \beta \cdot \text{StyleSim}(x_i, s_u) + \gamma \cdot \text{PrefScore}(x_i) \tag{11}$$

where CLIPScore: semantic consistency; StyleSim: style match (with user preference vectors); PrefScore: from the rating prediction model. StyleSim is implemented as the cosine similarity between CLIP image embeddings of the generated output and the style reference. This embedding-based metric captures both visual appearance and higher-level stylistic cues, offering a more semantically grounded measure than handcrafted features or Gram-matrix correlations. Then it is selected, poorer as well as mutated accordingly to generate a new population for the next generation as follows:

$$\mathcal{P}_{t+1} = \text{Generate}\left(\mathcal{P}_t^{\text{selected}}, \text{Crossover}, \text{Mutation}\right). \tag{12}$$

Keep iterating the algorithm until the corresponding requirement is finally satisfied. In this article, the overall structure flow of EAGAN, which is finally built by integrating the above text image generation, style transfer and adaptive iterative optimization, is shown in Fig. 2.

The proposed network, driven by user-provided textual input, integrates a diffusion-based generation mechanism, style transfer technology, and a genetic optimization strategy to construct a multi-module collaborative framework for the personalized creation of cultural and creative products. The process begins with the user articulating their creative intent *via* natural language, which is encoded into semantic vectors by the text encoding module. These vectors are then fed into a Stable Diffusion-based architecture, initiating image synthesis within a compact latent space. Simultaneously, user-supplied style references—whether in the form of exemplar images

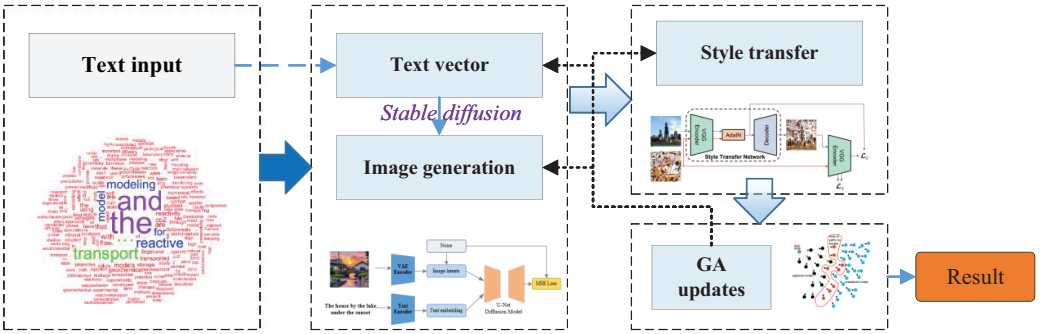

Figure 2 **The framework for the evolutionary adaptive generative aesthetic network.**
Full-size DOI: 10.7717/peerj-cs.3288/fig-2

or abstract preference vectors—are incorporated through the Style transfer Module, which employs the AdaIN mechanism to modulate feature representations and infuse stylistic attributes. Following initial image generation, the system conducts an evaluation of the output using a multidimensional fitness function that accounts for semantic alignment, style coherence, and subjective aesthetic congruence. Based on this assessment, a genetic algorithm iteratively refines the textual prompts, style vectors, and latent noise parameters, enabling adaptive optimization across successive generations. Through multiple evolutionary cycles, the system progressively converges on outputs that exhibit high semantic fidelity, distinctive stylistic traits, and strong personalization. To achieve multimodal alignment, both textual and visual features are projected into a shared embedding space. Specifically, we adopt the CLIP encoder (ViT-L/14 variant), which jointly encodes text and images into semantically aligned vector representations through contrastive pretraining. Text prompts and user preference vectors are processed by the CLIP text encoder, while images and generated outputs are embedded using the CLIP vision encoder. Semantic consistency is enforced by measuring cosine similarity within this shared space, which also serves as the basis for optimization in the Evolutionary Module. This design ensures that the generated imagery aligns not only with the textual description but also with the encoded stylistic and preference cues, providing a robust multimodal bridge for both training and evaluation. Abstract concepts such as tension, fantasy, or urgency are not modeled through explicit rules but are implicitly captured by the CLIP text encoder, which aligns them with recurring visual patterns in large-scale training data. These embeddings guide the generator toward corresponding stylistic or compositional features, though the mapping remains indirect and may introduce ambiguity in certain cases.

This comprehensive framework embodies a synergistic and iterative co-evolution across four key dimensions—text, image, style, and optimization—offering a robust, scalable, and user-adaptive solution for creative ideation and customized content generation.

The proposed model, EAGAN, is built upon the Stable Diffusion architecture and designed to enable personalized image generation guided by semantic and stylistic preferences. The input data consisted of paired text prompts and cultural image content,

sourced from benchmark datasets relevant to cultural and creative product design. During data preprocessing, all images were resized to a consistent resolution, and the corresponding text prompts were tokenized and encoded using a pre-trained semantic text encoder. Image normalization was applied to ensure consistency across visual inputs. Style references were also collected and standardized to feed into the AdaIN-based modulation module for effective style conditioning. The evolutionary component required encoding user preferences into cue phrases and latent vectors, which were initialized randomly and then iteratively evolved through a genetic algorithm.

Evaluation method

To assess the effectiveness of EAGAN, a comprehensive evaluation strategy was employed. Comparative analysis was conducted against state-of-the-art generative models, including baseline diffusion-based and transformer-based architectures. Ablation experiments were performed to isolate the individual contributions of the adaptive style modulation and evolutionary optimization modules. In these experiments, key components were selectively removed to observe performance degradation. Application-based evaluation was also conducted to determine the model's practical utility in real-world design scenarios by generating culturally inspired product concepts and assessing stylistic relevance and personalization through expert review panels. This multifaceted evaluation approach helped establish both the technical robustness and creative applicability of the model.

Assessment metrics

The performance of EAGAN was measured using a suite of quantitative metrics. Fréchet inception distance (FID) was used to evaluate the visual quality and realism of generated images by measuring the distributional distance between generated and real image features. CLIPScore was employed to assess semantic alignment between generated images and their corresponding text prompts, reflecting how accurately the model captures textual intent. Learned Perceptual Image Patch Similarity (LPIPS) was used to evaluate perceptual differences between generated outputs and ground truth images, thereby indicating the fidelity of style transfer and fine detail retention. These metrics collectively provided a balanced evaluation of semantic accuracy, visual quality, and stylistic coherence.

EXPERIMENT RESULT AND ANALYSIS

Dataset and experiment setup

Given that this study encompasses both text-to-image generation and style transfer, we employ subsets from the WikiArt, BAM, and LAION datasets to support comprehensive evaluation. The WikiArt dataset comprises over 80,000 paintings produced by artists worldwide, encompassing more than 20 canonical artistic styles such as Impressionism, Cubism, and Realism, and includes rich metadata on styles and creators. For this research, we selected 100 representative images from each of the ten most prominent styles, constructing a 1,000-image subset to assess the model's style control capabilities (*Eichler, Eichler & Del Pino, 2023*). In contrast, the BAM dataset (*Froehlich & Koeppl, 2024*) emphasizes modern design and illustration, featuring approximately 65,000 art images with clearly annotated style labels. Its categories—such as digital painting, graphic design,

and hand-drawn line art—align more closely with the practical aesthetic requirements of the cultural and creative industries. Similarly, we extracted 100 samples from each of ten styles within BAM and integrated them with the WikiArt subset to establish a consolidated benchmark dataset for evaluating style adaptability and control.

In order to evaluate the model's performance on text–image semantic consistency, we further utilized the LAION-5B dataset (*Schuhmann et al., 2022*). From this large-scale collection, we extracted a dedicated evaluation subset consisting of 1,000 high-quality image–text pairs. The curation process combined keyword-based filtering with manual inspection: candidate samples were first selected through thematic keyword queries, and then carefully reviewed to ensure that the associated textual descriptions not only captured the central thematic content but also included stylistic indicators. To maintain consistency and reduce noise across samples, several preprocessing steps were applied. All images were standardized to a fixed resolution of 512 × 512 pixels, and samples exhibiting low visual quality, vague or incomplete textual descriptions, or clear semantic mismatches between text and image were systematically excluded. This multi-stage filtering procedure ensured that only coherent and representative pairs were retained.

The final evaluation set was structured into two complementary subsets:

Stylistic expressiveness subset (sourced from WikiArt and BAM), focusing on artistic attributes, visual style, and expressive qualities.

Semantic alignment subset (sourced from LAION), emphasizing accurate correspondence and fidelity between textual descriptions and visual content.

Together, these curated subsets enable a robust, multi-dimensional evaluation protocol. By balancing stylistic control with semantic fidelity, the dataset supports comprehensive validation and comparative analysis of the proposed model's performance. Importantly, compared with the raw large-scale datasets, the curated subsets underwent stricter quality control and noise reduction, thereby enhancing the reliability and stability of the evaluation outcomes.

To ensure reproducibility and minimize evaluation bias, the dataset was divided into training and validation subsets following an 80/20 split, stratified across both the stylistic (WikiArt/BAM) and semantic (LAION) partitions to preserve representativeness. The validation set was kept entirely disjoint from training data to avoid information leakage. For user preference modeling, stylistic indicators were encoded using categorical embeddings, while semantic alignment cues were represented by normalized dense vectors derived from CLIP text embeddings; these preference vectors served as control signals during model conditioning. All experiments were conducted with fixed random seeds to guarantee reproducibility, and repeated runs yielded stable results with minimal variance. To mitigate potential evaluation bias, sample selection was balanced across categories, manual curation was cross-validated by multiple annotators, and reported metrics represent averaged performance across subsets rather than relying on isolated cases.

To comprehensively evaluate the performance of the proposed model in cultural and creative image generation, this study constructs a multidimensional evaluation framework encompassing four key dimensions: image quality, semantic consistency, style control, and personalized adaptability.

For image quality, two widely adopted metrics are employed: FID and Inception Score (IS). FID quantifies the distributional distance between generated and real images in the feature space, with lower values indicating higher realism. IS simultaneously considers image diversity and clarity, offering a complementary perspective on generative fidelity.

In terms of semantic consistency, the study uses CLIPScore, which computes the cosine similarity between the CLIP-encoded embeddings of the input text and the generated image. This metric reflects the model's ability to accurately capture and visually represent user intent as expressed through natural language input.

To assess style control ability, LPIPS distance and Gram Matrix similarity are employed. LPIPS evaluates perceptual differences in visual style, while Gram Matrix similarity captures correlations of feature activations related to texture, brushstroke, and color. Since style transfer tasks preserve semantic structure and only modulate visual appearance, IS is excluded from this component of the evaluation.

For personalized adaptability, the model's performance is gauged by its responsiveness to varying user preferences through iterative optimization, reflected in improved scores across the above dimensions under evolving input constraints.

To benchmark the effectiveness of the proposed framework, a suite of representative models is selected for comparison:

- Stable Diffusion (*Du et al., 2023*): a high-fidelity, text-driven baseline with strong semantic preservation capabilities.
- VQGAN+CLIP (*Crowson et al., 2022*): a hybrid model known for its abstract, expressive generation through joint optimization of image and text embeddings.
- StyleGAN-NADA (*Gal et al., 2022*): a semantically guided style transfer model, well-suited for direct comparison with the proposed style transfer module.

These models encompass the prevailing paradigms in generative AI and provide diverse, challenging baselines for comparative analysis. The experimental environment used for training and evaluation—including hardware specifications, software versions, and configuration details—is summarized in Table 1.

The genetic algorithm was configured with a population size of 30, a mutation rate of 0.2, and a crossover rate of 0.5, employing a tournament selection strategy. The number of iterations was fixed at 20, as preliminary convergence analysis showed that performance improvements largely stabilized beyond this point. To assess the robustness of our results, all experiments were repeated with three random seeds, and we report the average performance along with the corresponding standard deviation. This procedure ensures that the improvements observed in FID, CLIPScore, LPIPS, and Style Similarity are not due to seed-specific effects. The results show stable trends across runs, with small variances, indicating that the gains of EAGAN over baselines are statistically reliable.

Method comparison and result analysis

Following the completion of model construction, we conducted comprehensive testing using the two constructed datasets. The performance of the proposed model was evaluated

Table 1 The experiment environment information.

Environment	Information
CPU	I5-10700F
GPUs	RTX 4060
Language	Python 3.10
Framework	Pytorch

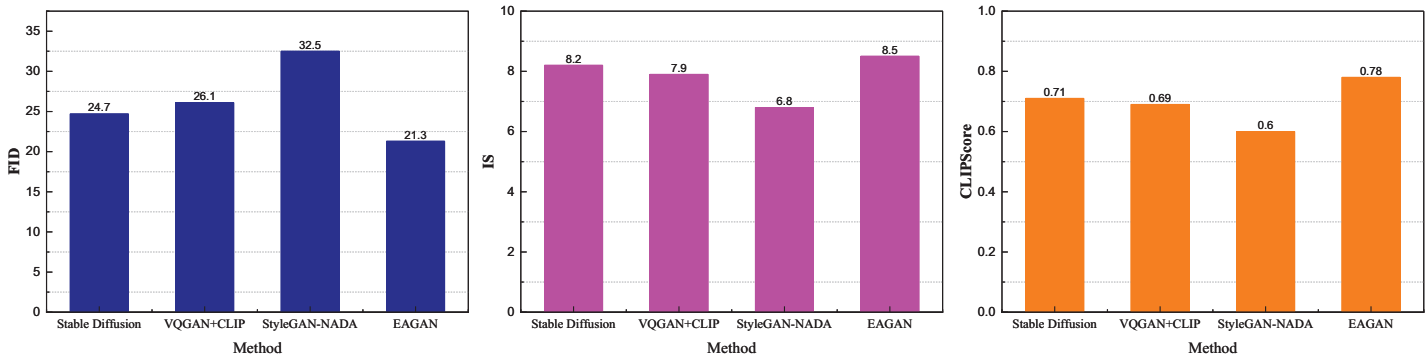

Figure 3 The method comparison results on LAION-5B datasets. Full-size DOI: 10.7717/peerj-cs.3288/fig-3

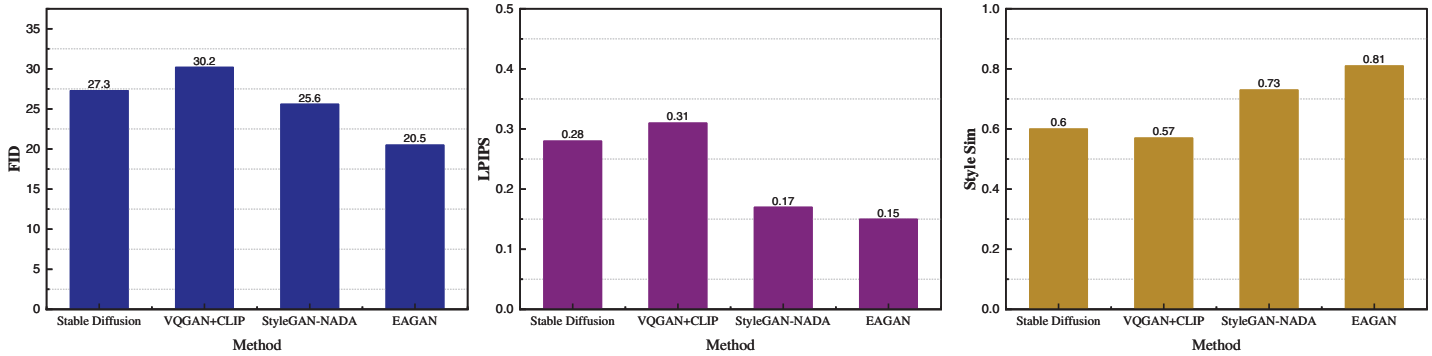

Figure 4 The method comparison results on BAM + WikiArt datasets. Full-size DOI: 10.7717/peerj-cs.3288/fig-4

across the primary metrics outlined earlier. The corresponding results—covering image quality, semantic consistency, and style control—are visually presented in Figs. 3 and 4, respectively, demonstrating the effectiveness of the model under varying input conditions and across multiple evaluation dimensions.

Figure 3 presents the comparative results of image generation performance across various methods on the LAION-5B dataset. Evaluated along three key metrics—FID, IS, and CLIPScore—the proposed EAGAN model demonstrates superior performance across all dimensions. Specifically, EAGAN achieves the lowest FID score (21.3), indicating that its generated images exhibit the highest resemblance to real image distributions. It also records the highest CLIPScore (0.78), evidencing strong semantic alignment between the

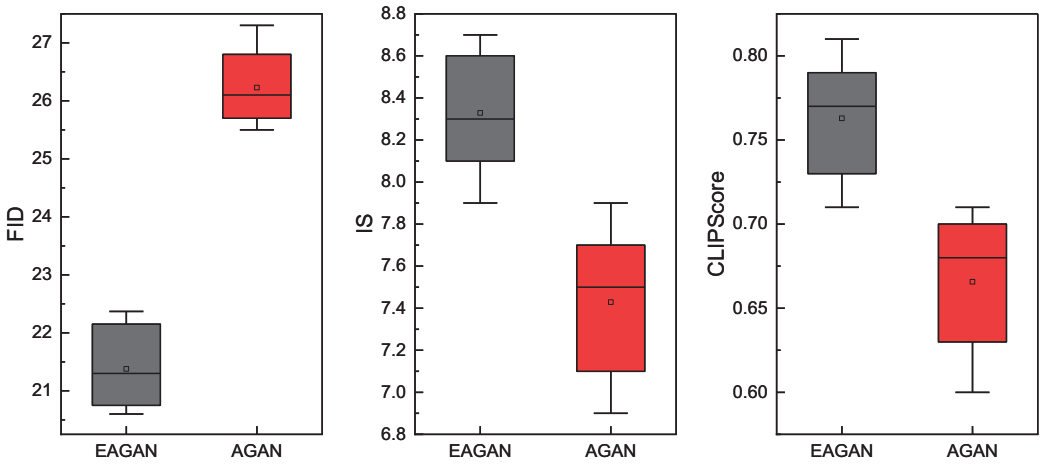

Figure 5 **The Ablation experiment results on LAION-5B datasets.**
Full-size DOI: 10.7717/peerj-cs.3288/fig-5

generated visuals and their corresponding textual descriptions. In terms of generative diversity and visual fidelity, EAGAN attains an IS of 8.5, outperforming baseline models and validating the effectiveness of its integrated evolutionary optimization and style regulation mechanisms.

Figure 4 illustrates the comparative performance of various methods in style transfer tasks on the BAM and WikiArt datasets. The proposed EAGAN model demonstrates consistently superior results across all evaluated metrics. Notably, EAGAN achieves the lowest FID score (20.5) and lowest LPIPS distance (0.15), indicating that its generated images maintain high visual quality and preserve intricate details while undergoing style transformation. In addition, the model attains the highest Style Similarity score (0.81), substantially outperforming StyleGAN-NADA (0.73) and other baseline approaches. These results highlight EAGAN's enhanced capacity for style expression and alignment, effectively validating the impact of its evolutionary optimization strategy and adaptive style regulation module.

Ablation experiment and application analysis

After benchmarking EAGAN against existing models, we further evaluated its evolutionary optimization performance, specifically the impact of the GA module, through a series of ablation experiments. The results of these analyses are presented in Figs. 5 and 6, which highlight the performance differences with and without GA integration. These comparisons reveal the critical role of the evolutionary component in enhancing semantic fidelity, style precision, and overall image quality, thereby confirming its effectiveness in driving adaptive and personalized generation.

Figure 5 demonstrates the impact of the evolutionary optimization module on model performance, with AGAN representing the ablation version of the model without GA integration. The comparison clearly shows that the full EAGAN model surpasses AGAN across all three key metrics—FID, IS, and CLIPScore. In particular, EAGAN achieves a

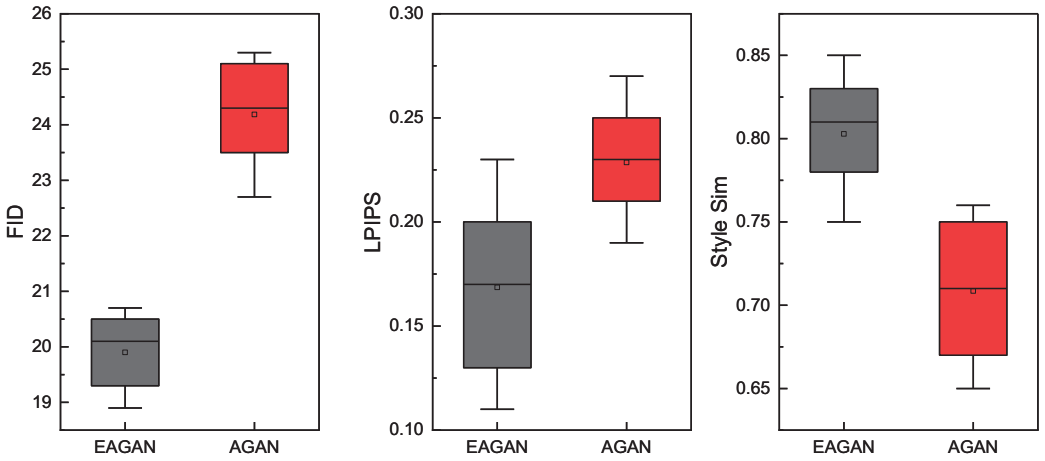

Figure 6 The ablation experiment results on BAM + WikiArt datasets.
Full-size DOI: 10.7717/peerj-cs.3288/fig-6

notably lower FID, indicating a substantial reduction in the distributional gap between generated and real images. Simultaneously, the higher CLIPScore reflects enhanced semantic alignment between textual prompts and generated visuals. These improvements validate the effectiveness of the GA module in adaptively exploring and refining the parameter space, thus reinforcing its crucial role in optimizing image quality and semantic fidelity.

Figure 6 presents the results of the ablation experiments focused on the style transfer task using the BAM and WikiArt datasets. When compared to AGAN—the variant lacking the evolutionary optimization module—the complete EAGAN model demonstrates superior performance across all evaluated metrics. Specifically, EAGAN records lower FID and LPIPS values, indicating the generation of higher-quality images with smoother and more coherent style transitions. Furthermore, its Style Similarity score approaches 0.85, significantly outperforming AGAN and underscoring the effectiveness of the genetic algorithm in enhancing style alignment. These findings affirm the pivotal role of the evolutionary optimization module in strengthening both the consistency of stylistic expression and the overall visual fidelity of the generated results.

Figure 7 illustrates the impact of varying the number of GA iteration rounds on model performance across both datasets. The results reveal that as the number of GA iterations increases, EAGAN consistently improves in terms of FID and LPIPS, reflecting enhanced image quality and better style fidelity through evolutionary refinement. On the LAION-5B dataset, both CLIPScore and IS reach their peak between 15 and 25 iterations, after which a slight decline suggests the emergence of a search equilibrium. In contrast, on the BAM + WikiArt datasets, Style Similarity continues to rise steadily with more iterations until it stabilizes, indicating a progressive enhancement in stylistic alignment.

Overall, these findings highlight the significant influence of GA iteration count on generative performance. Based on the observed trends, the model achieves its optimal

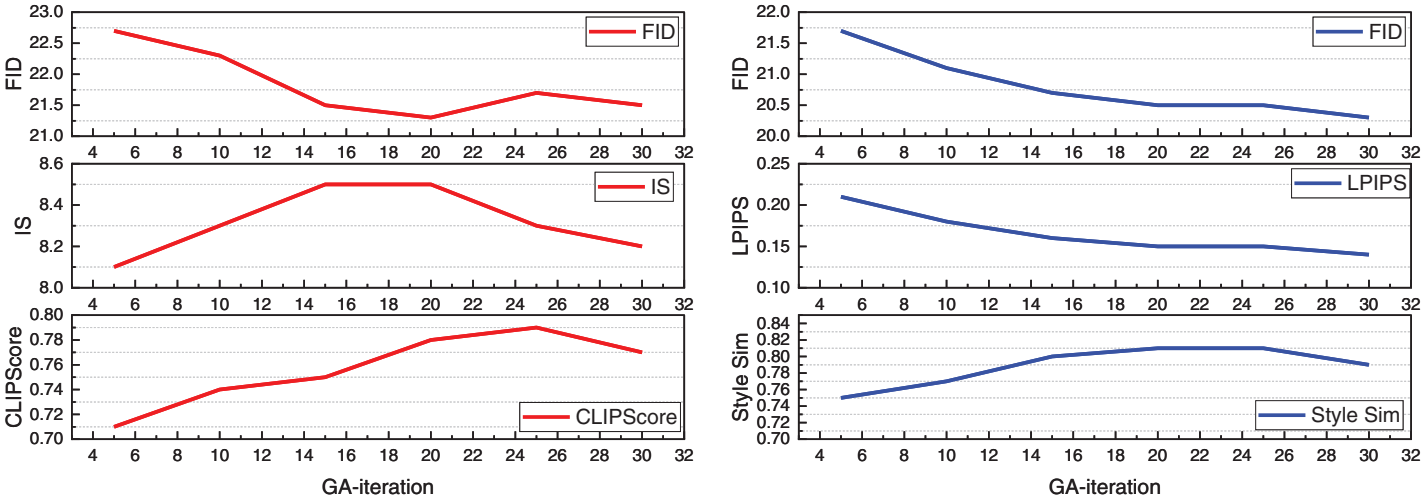

Figure 7 **The results for different GA iterations on both datasets.** Full-size DOI: 10.7717/peerj-cs.3288/fig-7

multidimensional results at 20 iterations, which is selected as the default configuration for this study.

Following the quantitative analysis, subjective human perception—an essential factor in evaluating personalized generative models—was assessed through a user study. Volunteers were asked to rate, on a 10-point scale, the visual quality and stylistic relevance of text-based images generated under varying GA iteration counts. As shown in Fig. 8, user scores exhibit a clear upward trajectory as the number of GA iterations increases, followed by a plateau. At the 5th iteration, the outputs were relatively unrefined, yielding an average score of 7.7, suggesting that early-stage evolutionary optimization was insufficient. However, between the 15th and 20th iterations, marked improvements in semantic accuracy and stylistic expression were observed, with the 20th iteration achieving the peak average score of 8.8. Beyond this point, user ratings fluctuate slightly but remain consistently high, indicating convergence in perceived quality. These results confirm that the GA module significantly enhances the aesthetic appeal, style alignment, and personalization of generated images. Moreover, the convergence trend supports the conclusion that a moderate iteration count effectively balances visual quality with computational efficiency for practical deployment scenarios.

DISCUSSION

The EAGAN proposed in this study integrates four key components—text semantic understanding, image generation, style transfer, and evolutionary optimization—to construct a highly collaborative framework for personalized image synthesis. In contrast to conventional text-to-image generation models such as Stable Diffusion or VQGAN+CLIP, EAGAN introduces two novel and synergistic modules: Style Adaptive Module and the Evolutionary Optimization Module, which significantly enhance its flexibility and generative precision. The Style Adaptive Module, built on the AdaIN mechanism, enables dynamic modulation of stylistic attributes by incorporating user-supplied style references

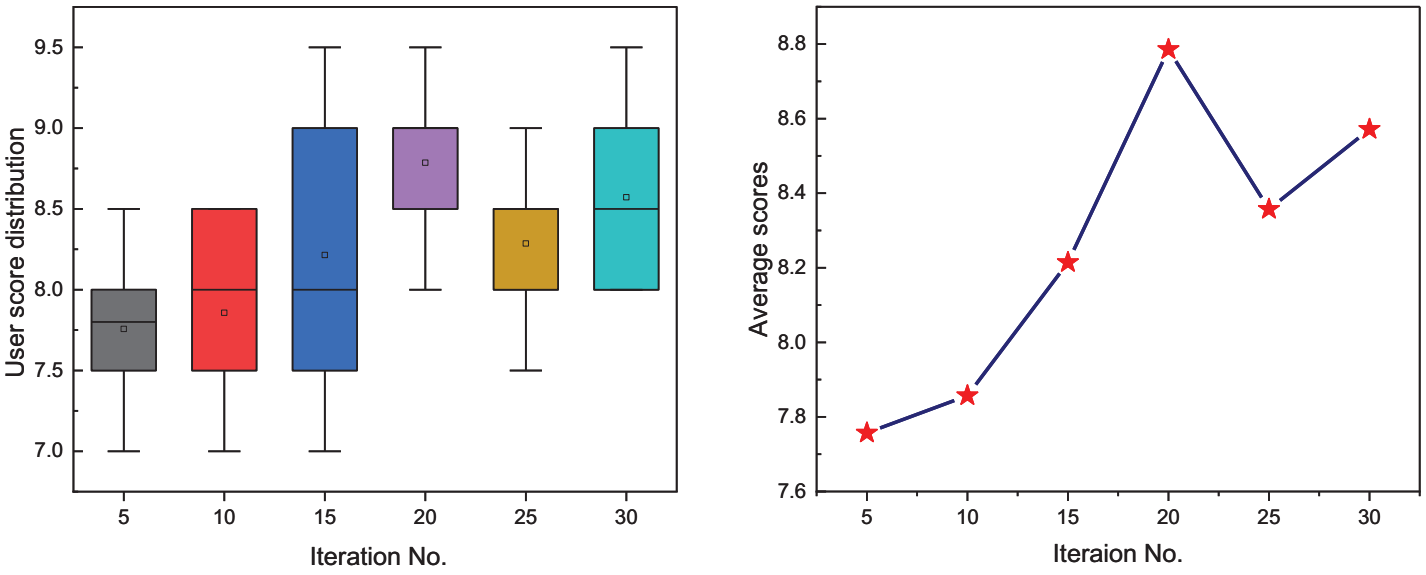

Figure 8 **The user evaluation scores for different GA iteration results.** Full-size DOI: 10.7717/peerj-cs.3288/fig-8

or preference vectors. This allows the model to maintain semantic integrity while achieving customized stylistic expression, effectively addressing diverse aesthetic expectations. Simultaneously, the Evolutionary Optimization Module employs a genetic algorithm to perform a global search and multi-round iterative refinement over prompt representations, style parameters, and latent noise vectors. This enables continuous enhancement of image quality, semantic alignment, and personalized fit—far surpassing the capabilities of traditional models constrained by static parameter sets. By coupling self-adaptive style regulation with evolution-driven optimization, EAGAN overcomes key limitations in controllability and expressiveness faced by existing approaches. Experimental evaluations across multiple benchmarks consistently demonstrate EAGAN's superior performance in FID, CLIPScore, and Style Similarity—especially in tasks demanding high fidelity in semantic consistency and style transfer—underscoring the robustness, adaptability, and innovation of its structural design. Although our evaluation highlights improvements in image fidelity and semantic alignment over baseline models, we acknowledge that the comparison does not systematically assess controllability of stylistic attributes or responsiveness to user preferences. These dimensions are critical for personalized generation but remain difficult to benchmark due to the absence of standardized protocols. While qualitative results suggest that EAGAN provides finer style modulation and more adaptive preference alignment than prior approaches, future work will focus on developing structured evaluation methods—potentially combining user studies with automatic metrics—to more rigorously validate these advantages.

The EAGAN model demonstrates substantial application potential in the personalized design of cultural and creative products, addressing the growing demand for user-driven customization as the industry shifts away from standardized production. In an era where creative output must integrate cultural depth, distinctive visual style, and individual

expression, EAGAN offers an intelligent, automated solution that seamlessly translates natural language input and user preferences into richly styled, semantically meaningful imagery. By fusing text semantics with adaptive style modulation, EAGAN enables the efficient generation of culturally resonant and aesthetically refined content. This capability is particularly valuable in scenarios such as digital collectibles, themed illustrations, stationery and packaging design, fashion pattern generation, and more. For instance, a designer or casual user can simply input prompts like "Dunhuang-style flying apsara" or "ink-style cat", and the system will produce visually compelling, stylistically aligned artwork—effectively lowering the creative entry barrier. Moreover, the GA-based evolutionary module allows for real-time, iterative refinement based on user feedback, enabling outputs to dynamically converge toward personalized aesthetic preferences. This interactive loop enables a genuine "human–machine co-creation" experience, in which creativity is not fully automated but instead adaptively guided by user input and iteratively refined through the model's feedback. Interactivity is modeled through a learned preference-adaptive mechanism: user adjustments to prompts, style vectors, or reference images are encoded and dynamically integrated into the generation process, allowing the system to respond in real time to evolving intentions. For animation or scene dynamics, the framework currently employs a hybrid strategy—rule-based temporal interpolation for smooth visual transitions, combined with learned feature modulation to maintain semantic and stylistic consistency across frames.

Beyond individual use, EAGAN can be embedded into cultural and creative platforms or interactive design tools, supporting applications such as assisted concept development, customized design generation, and cultural content reinterpretation. In this way, it not only improves design efficiency but also broadens participation in creative processes by providing an intelligent, personalized, and adaptive design partner. In the current implementation, our framework does not explicitly incorporate a visual attention mechanism. Instead, semantic–visual alignment is primarily achieved through the use of pretrained CLIP encoders and the evolutionary optimization strategy, which iteratively adjusts prompts, style vectors, and latent codes to improve coherence between text and image features. This design choice reflects our focus on preference-driven optimization and style adaptivity, rather than token-level attention modeling. That said, we acknowledge that integrating visual attention modules could further strengthen fine-grained text–image correspondence, particularly in complex scenes where localized alignment is critical. For example, cross-attention mechanisms between text tokens and image regions, as employed in recent diffusion-based models, may enhance controllability over object placement and region-specific styling. Exploring the integration of such attention-based strategies into our framework represents an important direction for future work.

CONCLUSION

In this article, we introduced EAGAN, a novel framework for personalized cultural and creative image generation that integrates semantic alignment, style adaptivity, and preference-aware optimization. By combining the generation strengths of Stable Diffusion,

the flexible transformation capabilities of AdaIN, and the adaptive search properties of evolutionary algorithms, EAGAN enables a unified system for controllable and user-responsive image synthesis. The proposed Style Adaptive Module effectively injects stylistic cues from user-provided references or encoded preference vectors, while the Evolutionary Optimization Module iteratively refines prompt formulations and latent parameters to enhance semantic coherence and personalization. Experimental results on datasets such as LAION-5B, BAM, and WikiArt show consistent improvements over baseline methods (Stable Diffusion, VQGAN+CLIP, and StyleGAN-NADA) on widely adopted metrics including FID, CLIPScore, LPIPS, and Style Similarity. Ablation studies further highlight the contribution of both the style control mechanism and the evolutionary optimization strategy to these gains.

At the same time, we acknowledge several limitations of the present study. The user study design and participant selection were limited in scale and diversity, constraining the generalizability of subjective evaluations. Moreover, while benchmark results demonstrate promising performance, further in-depth error analysis and statistical validation are needed to more rigorously establish significance. Finally, although EAGAN shows strong adaptability, broader testing across additional cultural domains and creative contexts remains necessary.

Looking ahead, future work will focus on conducting larger and more systematic user studies, incorporating robust statistical analysis, and extending the framework to encompass richer modalities of user interaction and cultural specificity. By addressing these directions, we aim to further strengthen the reliability and applicability of EAGAN in supporting personalized, culturally grounded visual content creation.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers whose comments and suggestions helped to improve the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Dailei Hu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Enshi Wang performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.

- Muddassira Arshad analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The WikiArt is available at Hugging Face and Zenodo:

- https://huggingface.co/datasets/huggan/wikiart.

- Lisi, E., Malekzadeh, M., Haddadi, H., Lau, F. D.-H., & Flaxman, S. (2020). Modeling and forecasting art movements with CGANs [Data set]. Zenodo. https://doi.org/10.5061/dryad.90cj2pq.

The BAM data is available at GitHub: https://github.com/google-research-datasets/bam.

The LAION dataset is available at: https://laion.ai/laion-5b-a-new-era-of-open-large-scale-multi-modal-datasets.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3288#supplemental-information.

REFERENCES

Bermano AH, Gal R, Alaluf Y, Mokady R, Nitzan Y, Tov O, Patashnik O, Cohen-Or D. 2022. State-of-the-art in the architecture, methods and applications of StyleGAN. *Computer Graphics Forum* **41(2)**:591–611 DOI 10.1111/cgf.14503.

Cao K, Liao J, Yuan L. 2020. CariGANs: unpaired photo-to-caricature translation. *ACM Transactions on Graphics* **37(6)**:244.1–244.14 DOI 10.1145/3272127.3275046.

Cheng J, Wu FX, Tian YL, Wang L, Tao DP. 2020. RiFeGAN: rich feature generation for text-to-image synthesis from prior knowledge. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 10911–10920.

Crowson K, Biderman S, Kornis D, Stander D, Hallahan E, Castricato L, Raff E. 2022. VQGAN-CLIP: open domain image generation and editing with natural language guidance. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 88–105.

Du C, Li Y, Qiu Z, Xu C. 2023. Stable diffusion is unstable. *Advances in Neural Information Processing Systems* **36**:58648–58669.

Eichler TZN, Eichler ML, Del Pino JC. 2023. Serendipidade e Arte na Educação em Química: Apresentando a WikiArt, uma Enciclopédia de Artes Visuais. *Revista Debates em Ensino de Química* **9(4)**:92–106 DOI 10.53003/redequim.v9i4.4902.

Froehlich P, Koeppl H. 2024. Graph structure inference with BAM: neural dependency processing via bilinear attention. In: *Advances in Neural Information Processing Systems 37*.

Gal R, Patashnik O, Maron H, Bermano AH, Chechik G, Cohen-Or D. 2022. StyleGAN-NADA: clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* **41(4)**:1–13 DOI 10.1145/3528223.3530164.

Gu J, Ye JC. 2021. AdaIN-based tunable CycleGAN for efficient unsupervised low-dose CT denoising. *IEEE Transactions on Computational Imaging* **7**:73–85 DOI 10.1109/tci.2021.3050266.

Han S, Shi Z, Shi Y. 2022. Cultural and creative product design and image recognition based on the convolutional neural network model. *Computational Intelligence and Neuroscience* **2022(1)**:2586042 DOI 10.1155/2022/2586042.

Hou H, Huo J, Wu J, Lai YK, Gao Y. 2021. MW-GAN: multi-warping GAN for caricature generation with multi-style geometric exaggeration. *IEEE Transactions on Image Processing* **30**:8644–8657 DOI 10.1109/tip.2021.3118984.

Hui C. 2021. Research of user-centered intelligent technology in China's cultural and creative product design. In: *E3S Web of Conference*. Vol. 236, EDP Sciences, 4050.

Jang W, Ju G, Jung Y, Yang J, Tong X, Lee S. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *ACM Transactions on Graphics (TOG)* **40(4)**:1–16 DOI 10.1145/3450626.3459860.

Li R, Wang C. 2022. Cultural and creative product design and image recognition based on deep learning. *Computational Intelligence and Neuroscience* **2022(1)**:7256584 DOI 10.1155/2022/7256584.

Liao WT, Hu K, Yang MY, Rosenhahn B. 2022. Text to image generation with semantic-spatial aware GAN. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 18187–18196.

Men Y, Yao Y, Cui M, Lian Z, Xie X, Hua XS. 2022. Unpaired cartoon image synthesis via gated cycle mapping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 3501–3510.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. 2021. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. Westminster: PMLR, 8748–8763.

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In: *Advances in Neural Information Processing Systems 35*.

Tan ZR, Yang X, Ye ZH, Wang QF, Yan YY, Nguyen A, Huang KZ. 2023. Semantic similarity distance: towards better text-image consistency metric in text-to-image generation. *Pattern Recognition* **144**:109883 DOI 10.1016/j.patcog.2023.109883.

Tao M, Tang H, Wu F, Jing XY, Bao BK, Xu CS. 2022. DF-GAN: a simple and effective baseline for text-to-image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 16515–16525.

Tominaga R, Seo M. 2022. Image generation from text using StackGAN with improved conditional consistency regularization. *Sensors* **23(1)**:249 DOI 10.3390/s23010249.

Turja SD, Jubair MI, Rahman MS, Al Zadid MH, Shovon MH, Khan MFK. 2022. Shapes2Toon: generating cartoon characters from simple geometric shapes. In: *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*. Piscataway: IEEE, 1–8.

Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. 2020. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 1316–1324.

Ye SM, Wang H, Tan MK, Liu F. 2024. Recurrent affine transformation for text-to-image synthesis. *IEEE Transactions on Multimedia* **26(14)**:462–473 DOI 10.1109/tmm.2023.3266607.

Zhang J. 2021. Exploring the application of traditional elements in cultural and creative product design. *Art and Design Review* **9(4)**:332–340 DOI 10.4236/adr.2021.94029.

Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. 2019. StackGAN++: realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41(8)**:1947–1962 DOI 10.1109/tpami.2018.2856256.