

# Application of AI Technology in Art Concept Map Based on Intelligent Information Technology

Miao Hu

School of Fine Arts and Design, Mudanjiang Normal University, Mudanjiang 157012, Heilongjiang, China

Corresponding author: Miao Hu (e-mail: m13514565983@163.com)

**ABSTRACT** This study presents a dual-path collaborative generation framework for art concept maps based on multimodal inputs, including textual descriptions, hand-drawn sketches, and style reference images. The system integrates a precise control flow, leveraging ControlNet for structural and semantic alignment, with a creative inspiration flow guided by a style memory module to enhance artistic expression. Both streams are adaptively fused via a gated network, enabling dynamic weighting across spatial regions and generation stages. Experimental validation on a curated high-resolution concept art dataset demonstrates that the proposed framework achieves a Fréchet Inception Distance (FID) of 18.73 and a CLIP score of 0.321, outperforming baseline models including Stable Diffusion + ControlNet (FID 24.80, CLIP 0.293), Midjourney (FID 20.10), and LoRA-based variants (FID 23.50). Subjective evaluation by 150 professional concept artists confirms superior semantic accuracy, structural controllability, and stylistic consistency. By interpreting the dual-path framework as a multi-node signal acquisition and propagation system with dynamic fusion control, the study provides an engineering-oriented methodology for multimodal information fusion, precision guidance, and high-fidelity generation of structured visual content.

**INDEX TERMS** dual-path framework, multimodal input, gated fusion, structured generation

## I. INTRODUCTION

AS the digital creative industry continues to grow rapidly, the need for artwork to create art concept drawings for the games, film/TV, animation, and VR industries continues to increase and has led to increasing demand for shorter production cycles and higher output. Concept artwork created in the early design phase must accurately represent the characters, the atmosphere of the scene, and the overall narrative logic, while conveying a high level of artistic expression [1-2]. Similar requirements are increasingly prevalent in the modern textile and fashion industries where designers must rapidly iterate conceptual patterns while maintaining the precise structural logic of fabrics. This creates a two-fold challenge for creators in terms of their ability to create and produce work efficiently [3]. In digital textile design, the manual process of balancing intricate motif layouts with diverse artistic styles remains a significant bottleneck. Moreover, traditional manual drawing methods are labor-intensive and slow, making it impossible to fulfill the industry's immediate needs for efficiency and high quality [4]. Among the many breakthroughs provided by the emergence of intelligent information technology (such as educational and/or generative AI), one of the greatest advantages provided by generative AI and deep

learning via massive datasets is the ability to quickly create a broad array of visual content in multiple formats (including text descriptions and sketches) that can be generated from an extensive photo library, thereby providing significantly improved efficiency and scope of creative exploration [5,6]. In essence, models such as Stable Diffusion have enabled the evolution of text-to-image generation, reducing the burden of repetitive image creation and allowing artists to focus on their artistic vision and development [7,8]. Technology has transformed the way artwork is created [9,10], facilitating a shift from purely human-produced art to a new collaborative relationship between humans and machines. In other words, it creates more opportunities to solve the conflict between the efficiency and the quality of creation.

The landscape of art design has recently been impacted significantly by the use of intelligent information technology. Whereas systems such as computer-aided design were previously used as passive tools that helped designers produce higher-quality illustrations and modify them more efficiently, recent advances in AI-generated content have transformed this paradigm. Rather than merely complementing human creativity, technologies based on generative adversarial networks and variational autoencoders now enable designers to automatically incorporate artistic

expression into their creations [11-14]. The advent of significant technological breakthroughs in large-scale text-to-image generation such as those created with DALL-E [15], Stable Diffusion, and Midjourney [16] has allowed designers to create complex visual imagery with little-to-no human involvement through the use of easy-to-use interfaces and expansive creative options. Through the ability to generate imagery based solely on natural language cues, designers now have significantly increased possibilities for creative expression through the development of imagery and also explore completely new avenues of creativity. Control mechanisms, like ControlNet [17], can now direct the generation process using visual cues like sketches, positions, and outlines, allowing for greater control over the resulting output. At the same time, fine-tuning methods such as Low-Rank Adaptation (LoRA) [18,19] and DreamBooth [20] have been used effectively to reproduce specific artistic styles. These many technological advances have combined to move video design into a modern collaborative relationship between humans and machines, where artificial intelligence is no longer merely a tool, but rather serves as a creative ally, inspiring creativity and accelerating iteration on new work.

Research and practice to-date related to generative Artificial Intelligence's impact on art design has advanced greatly but has also highlighted a central contradiction with respect to creating automatically-generated artwork; there is still a large disparity between Artistically Creative Value versus Semantic Correctness across the various applications/implementations of this technology. Most studies/applications today utilize only one of these modes when developing their solution(s): (1) Controlling for compositional elements, proportions, and placement of design features will assure that created artwork meets all required parameters of the original design intent [21]; (2) reproducing a distinctive style to communicate messages or tell stories will frequently yield results that are artistically unique but do not accurately represent semantic content due to the variations inherent in stylized artwork created via the generative AI process [22,23]. Because of this separation, AI tools utilized by designers are not being incorporated effectively into the full extent of creative activities. As such, designers are unable to depend solely upon the use of AI for inspiration nor for generating usable designs. Studies have shown that in high-precision design tasks, more than 70% of professional artists believe that current AI output are "rich in creativity but uncontrollable" [24], or "structurally accurate but lacking artistic expression" [25], reflecting the deep conflict between controllability and inspiration. This contradiction essentially stems from the fragmented development of technological paths: the lack of a unified framework to integrate multimodal input and multi-level generation goals has resulted in human-computer collaboration remaining at a shallow level of interaction [26]. Therefore, it is urgent to build a new generation mechanism that can collaboratively optimize semantic precision control and artistic creative expression to achieve the goal of truly embedding AI into

the core creative process.

To address the core contradiction of current generative AI in balancing semantic accuracy and artistic creativity in the creation of concept art, this study proposes a dual-path collaborative generation framework based on Stable Diffusion that integrates multimodal conditional control and a style memory mechanism.

The integration of textual semantics, sketch structure, and stylistic reference images enables both precise controllability and expressive artistic generation. In this study, a dataset of high-quality concept artwork was constructed and systematically examines the generated results; performance was evaluated on 7 aspects: (1) overall semantic accuracy, (2) artistic creativity, (3) overall practical usability, (4) subjective evaluations of 150 professional artists; and (5) objective metrics: FID and CLIP score. The results demonstrate that the new methodology has produced an outstanding result on the FID of 18.73 and the CLIP score of 0.321, clearly exceeding the FID baselines in the generated artwork. The evaluation results of 150 professional concept artists who participated indicate that the new methodologies score better than the other mainstream models regarding usability and creative inspiration, thereby greatly increasing the visual diversity and stylistic expression whilst maintaining structural control, thus providing designers with additional tools for inspiring creative inspiration. Furthermore, while this study illustrates a pragmatic and data-driven approach to resolving the dilemma between control and creativity in AI-generated art, it also provides a foundation for understanding how intelligent information technology may empower art design through further methodological development.

## II. ART CONCEPT MAP AI GENERATION FRAMEWORK BASED ON INTELLIGENT INFORMATION TECHNOLOGY

### A. SYSTEM ARCHITECTURE

This paper proposes a dual-path collaborative generation framework based on the Diffusion Model, aiming to balance controllability and creativity. By running a precision control flow and a creative inspiration flow in parallel, the framework achieves an organic fusion of structural controllability and visual creativity.

Ultimately, a dynamic fusion module outputs high-quality conceptual images that combine professional usability with artistic expression. The overall framework is shown in Figure 1.

Using a multimodal input approach, the system supports three types of input: written descriptions, hand-drawn sketches, and reference images representing artistic styles. Text descriptions provide basic information about the artwork, including the identity of the character, emotional tone, and other contextual attributes. Hand-drawn sketches specify the spatial layout and relative positions of elements in the final composition. Reference images represent the overall artistic style of the artwork, including color palettes, brush-stroke textures, and lighting effects. The system integrates these three input modalities to generate the output, which

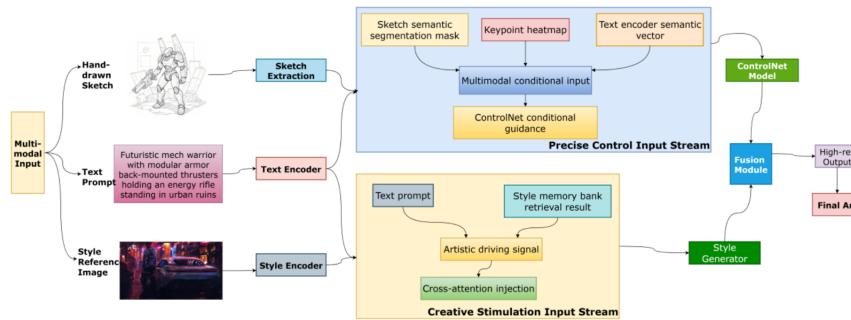


FIGURE 1. AI generation framework for art concept maps

is subsequently fed into the next stage of the dual-path generation process to produce the final results.

A precise control flow of image generation, invites sketch structure through the visual representation of text semantics, by pre-processing them with Edge Detection and Segmentation Techniques to extract contour and region division data. The result can then be injected as a spatial guidance signal into the diffusion model to guide layer-by-layer image generation during the denoising process using a ControlNet-like mechanism. Thus, during the denoising, images are controlled by their respective structures (i.e. character proportionality, 3D perspective, element layout) according to the design as defined in the original design documents. The combination of these elements creates an image generation process which prioritizes visual interpretability and logical accuracy making this process ideal for building core region generation tasks where strict compliance with design specifications is essential.

The Creative Inspiration Stream consists of multiple streams of creativity and is based on text prompts or style references. The model builds a style memory in which visual characteristics from a large collection of artworks from various sources are collected and stored. The model then selects from style memory the style vector which corresponds closest to the current task, and incorporates this into the image generation process of the latent space by applying various artistic treatment types in non-critical areas, including but not limited to background details, materials, lighting, etc. This path, while not being restricted to any specific type of forms, encourages designers to explore a variety of paths in order to produce visual inspiration they may not have been able to create initially.

The pathways used to generate artwork are not completely independent but are integrated into a unified module through a mechanism called a gated attention network. The contribution of each pathway to the final product is determined by a combination of user preference and task complexity. For instance, priority is given to the precise control flow in regions containing the main character to ensure accurate structural representation. Conversely, greater emphasis is placed on the creative inspiration flow when rendering background or atmospheric regions to enhance artistic expressiveness. Finally, the combined feature sets are processed and sent to a decoder to generate high-resolution,

well-organized artwork with a consistent visual style.

Designers can modify the input and fusion parameters to iteratively refine the generated artwork through multiple rounds of feedback and optimization.

## B. MULTIMODAL INPUT AND FEATURE EXTRACTION

The framework contains an integrated multimodal input processing system, which performs both semantic parsing and feature encoding, using the user's text prompts, sketches, and style reference images.

In this framework, the new deep semantic understanding technology utilizes the large-scale pre-trained T5 (Text-To-Text Transfer Transformer) language model to transform descriptive text input from the user into a semantic vector in high dimensionality. With the help of Contrastive Language-Image Pre-training (CLIP) a cross-modal encoder [27], deep semantic parsing will be performed on the text, and as well as on the traditional identification of keywords, adjectives, metaphorical language, and associative logical relationships associated with text input. The deep semantic parsing process converts natural language into a unified semantic space, creating a compact representation that describes an object by its object identity, physical attributes, ambient mood of an object, and emotional tone.

Given an input text  $T = \{t_1, t_2, \dots, t_n\}$ , where  $t_i$  is the word unit after word segmentation, refer to the research of Alaluf et al. [28] and map it into a high-dimensional semantic vector through the text encoder  $E_{\text{text}}$ :

$$v_t = E_{\text{text}}(T) \in \mathbb{R}^d \quad (1)$$

The vector  $v_t$  represents a number of abstract semantic properties, including object classification, material characteristics, environment ambience, and emotional tone, and serves as a common semantic reference point for 2-D and/or 3-D content generation by both paths. As an illustration, the phrase “metal armor (with flame patterns)” represents not only an explicit activation of the semantic node “armor” but also activates implicit characteristics such as “red/orange palette” and “glowing texture”.

Computer vision technology is applied to perform structured analysis on user-drawn line drafts. Original sketches usually contain rough outlines, key component locations, and perspective diagrams, but lack clear semantic labels

and geometric specifications. The system performs grayscale normalization and noise removal on the input image, uses the Canny edge detection algorithm to enhance line continuity, and uses a deep learning-driven semantic segmentation network [29] to identify functional categories in different regions.

The semantic segmentation network  $F_{\text{seg}}$  is applied to perform functional division on the sketch area, dividing the picture into logical areas such as the character body, background, and props. If the posture of a person or creature is involved, a lightweight posture estimation model is further used to extract the key point heat map  $H_k$  to form a set of spatial structure constraints:

$$M = F_{\text{seg}}(S), \quad H_k = F_{\text{pose}}(S) \quad (2)$$

$$C_{\text{structure}} = \{M, H_k\} \quad (3)$$

In formula (2),  $S$  is the hand-drawn sketch input by the user, usually a grayscale line drawing,  $F_{\text{seg}}$  is the semantic segmentation network, which outputs the category mask  $M$  to divide the character body, background, weapon and other areas;  $F_{\text{pose}}$  is the lightweight pose estimation model, which outputs the key point heat map  $H_k$ , which represents the probability distribution of joint positions.

$C_{\text{structure}}$  in formula (3) represents a set of structural control signals. These structural signals are used as spatial conditions for precise control flow. Composition lock is achieved by injecting the ControlNet mechanism into the diffusion model, ensuring that the generated image is geometrically consistent with the original sketch. A pose estimation algorithm is used to extract the character's key point skeleton, further restoring its dynamic pose and spatial orientation. The processed results are encoded into a spatial guidance graph, including edge maps, segmentation masks, and heat maps, which serve as the core input conditions for precise control flow.

Referring to the research of Chen et al. [30], a learnable style memory module is constructed as an adaptive style encoder to store and reuse the visual features of a specific artistic style. The style reference images uploaded by users (such as screenshots of an artist's works or classic concept images) are preprocessed and then extracted by a dedicated style encoder to extract multi-level artistic expression features. Given a style reference image, its non-content visual features are extracted by the adaptive style encoder. This process models the correlation between feature channels based on the Gram matrix. Assuming that the feature map of the  $l$ -th layer of the CLIP visual encoder is  $F_l \in \mathbb{R}^{C_l \times (H_l \times W_l)}$ , its style representation is defined as:

$$G_l = \frac{1}{C_l H_l W_l} F_l F_l^T \quad (4)$$

In formula (4),  $C_l$  is the number of channels in the feature map of that layer;  $H_l$  represents the height;  $W_l$  represents the width;  $F_l^T$  is the transpose of  $F_l$ ;  $G_l \in \mathbb{R}^{C_l \times C_l}$  is the

Gram matrix of layer  $l$ , reflecting the correlation between different feature channels; and  $C_l H_l W_l$  is the normalization factor used to eliminate scale differences and make features of different resolutions or depths comparable.

The final style vector  $v_s$  is obtained by concatenating and reducing the dimensionality of the Gram matrices of multiple layers. The calculation formula is as follows:

$$v_s = \text{MLP} \left( \bigoplus_{l \in L} G_l \right) \quad (5)$$

In formula (5),  $L$  represents the set of selected network layers used for multi-scale style modeling;  $\bigoplus$  represents the concatenation operation, which flattens the Gram matrices from different layers and concatenates them into a single long vector; Multi-Layer Perceptron (MLP) is a multi-layer perceptron used for dimensionality reduction and nonlinear transformation; and  $v_s \in \mathbb{R}^{d_s}$  is the final low-dimensional style vector output for storage and retrieval.

All style vectors, along with their metadata (such as style name and source tag), constitute a retrievable style memory. During generation, the system matches style vectors solely through non-structural aesthetic tags defined by the user prompt to ensure the retrieved artistic features augment rather than override the foundational creative intent. The construction and embedding process of the style memory is shown in Figure 2.

The encoding process of the style memory module targets non-structural aesthetic elements rather than specific content semantics. These elements entail color distribution, brushstroke texture, lighting treatment, and overall image atmosphere. All extracted style features are compressed into low-dimensional style vectors and stored in the style memory module along with metadata (such as style name, author, and application scenario).

During the generation phase, the system matches the most similar style vector based on the current task and injects it into the latent space or attention module of the diffusion model, achieving efficient transfer and reproduction of specific artistic styles. This mechanism allows designers to quickly switch style orientations. For example, it can generate multiple variations—such as cyberpunk or Chinese ink painting—for the same character setting, thereby greatly enhancing creative flexibility.

### C. DUAL-PATH COLLABORATIVE GENERATION MECHANISM

The deep fusion mechanism proposed in this paper is based on the Stable Diffusion Model. Its backbone network adopts a stable diffusion architecture and achieves high-quality image generation through iterative denoising in the latent space. This method uses a pre-trained U-Net as the core denoising network and incorporates a multimodal conditional control mechanism to support guidance from three modalities: text, structure, and style. A deep fusion mechanism is proposed within a dual-path architecture.

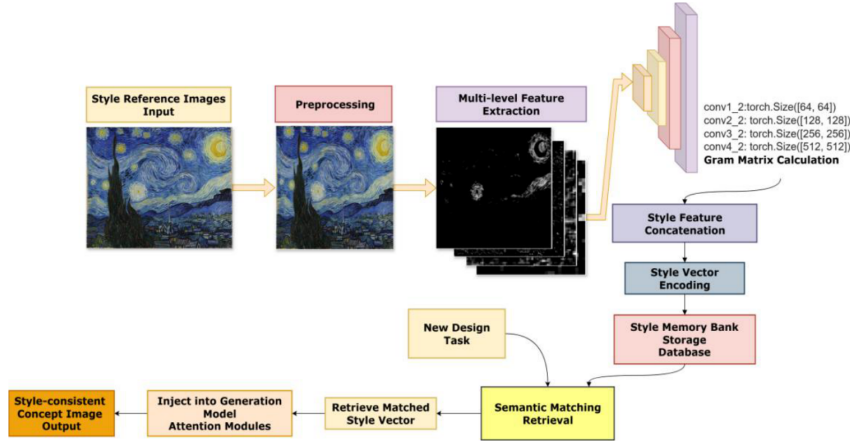


FIGURE 2. Construction and embedding of style memory module

This mechanism ensures faithful expression of design intent through a precise control flow and unleashes the artistic potential of the generative model through a creative inspiration flow. A dynamic fusion strategy is applied to achieve adaptive synergy between the two. This dynamic fusion strategy is illustrated in Figure 3.

Figure 3 illustrates the dynamic fusion strategy using Van Gogh's *The Starry Night* as an example.

The precise control flow receives the preprocessed sketch semantic segmentation mask  $M$  and keypoint heatmap  $H_k$ , combined with the semantic vector  $v_t$  output by the text encoder, as the multimodal conditional input to the diffusion model. Using the ControlNet conditional guidance mechanism, the sketch semantic segmentation mask and the keypoint heatmap serve as control conditions to inject spatial structure signals into each layer of the U-Net denoising process, ensuring that the generated image is strictly aligned with the user sketch in terms of layout, proportion, and element placement. In character design tasks, the system ensures that head-to-body ratio, limb orientation, and equipment distribution conform to the settings; in scene composition, it maintains the accuracy of perspective relationships and spatial hierarchy.

The creative inspiration flow uses textual prompts and search results  $v_s$  from the style memory library as primary driving signals, guiding the generative model to conduct diverse explorations in non-core areas. In areas that do not affect the main structure, such as background environment, material textures, lighting and shadow rendering, or atmospheric details, the model is encouraged to apply artistic brushstroke variations, color gradients, or abstract processing. The style vector  $v_s$  is injected into the self-attention layer of the diffusion model via a cross-attention mechanism. At the feature mapping stage, adaptive instance normalization is used to adjust channel statistics to achieve multi-level transfer of style features. In the intermediate feature layers of the U-Net (layers 3-6), the style vector  $v_s$  modulates the semantic-style association through attention weights, and adaptive instance normalization is used to align

the mean and variance of the feature maps, ensuring that the generated results both have the texture details and color tone of the target style.

After the two paths run in parallel, their potential representations need to be integrated through a dynamic fusion strategy. This paper proposes an adaptive fusion gate as a core innovation module. Based on the fusion mechanism of the gated network [31], it can adaptively adjust the contribution weights of the two outputs according to the task stage and regional importance. Let the feature map of the precise control flow be  $F_c \in \mathbb{R}^{C \times H \times W}$  and the feature map of the creative inspiration flow be  $F_e \in \mathbb{R}^{C \times H \times W}$ . The gated network splices the two and passes through the convolution layer and the Sigmoid activation function to generate a spatial weight map  $W \in [0, 1]^{H \times W}$ :

$$W = \sigma(\text{Conv}([F_c, F_e])) \quad (6)$$

In formula (6),  $[F_c, F_e]$  is the feature concatenation operation, which merges two feature maps along the channel dimension to obtain a tensor of size  $\mathbb{R}^{2C \times H \times W}$ ;  $\text{Conv}(\cdot)$  is a convolutional layer ( $3 \times 3$  convolution), which is used to learn spatial attention patterns from the concatenated features and output a single-channel feature map  $\mathbb{R}^{1 \times H \times W}$ ;  $\sigma(\cdot)$  is the sigmoid activation function, which compresses the output to the interval  $[0, 1]$  to ensure that the weight value of each spatial position is between 0 and 1; the resulting spatial gating weight map  $W$  represents the relative importance given to the precise control flow at each pixel position. The convolutional layers automatically identify core semantic regions by responding to high edge density from the sketch mask and semantic activation areas from the cross-attention maps.

The final fusion feature calculation formula is:

$$F_{\text{fused}} = W \odot F_c + (1 - W) \odot F_e \quad (7)$$

In formula (7),  $W$  is the spatial weight map from formula (6), with shape  $H \times W$  and broadcast to all channels;  $F_c$  is the control flow feature map,  $C \times H \times W$ ;  $F_e$  is the creative

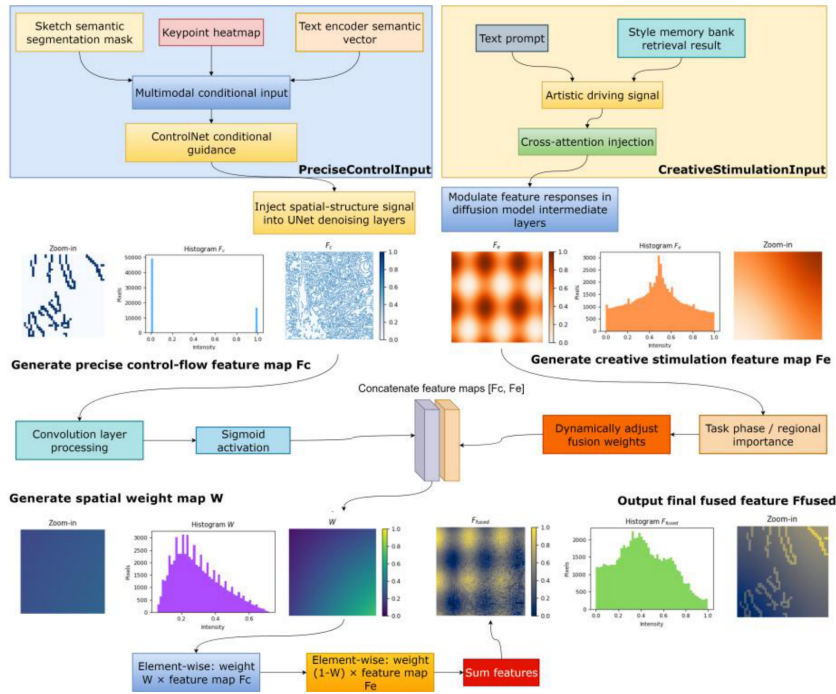


FIGURE 3. Dynamic fusion strategy based on gated network

flow feature map,  $C \times H \times W$ ; and  $F_{\text{fused}} \in \mathbb{R}^{C \times H \times W}$  is the final fused feature map, which serves as the input for the next stage of diffusion denoising. Element-wise multiplication  $\odot$  gives the ability to use weights based on pixel intensity. The gating functionality allows the system to apply a greater amount of weight to the  $F_c$  inputs for semantic regions (e.g., main characters, key props) to maintain structural stability, as well as applying greater amounts of weight to the  $F_e$  inputs for areas with considerable background and edge transitions for increased artistic variety. Dynamic control over fusion weights is also available through increasing the number of generation iterations. In the beginning, the emphasis is on generating creative ideas to inspire, while in later stages, the control of the flow of ideas becomes greater in order to reach an output that is usable.

Table 1 details the technical architecture of this system. The control flow structure integrates ControlNet technology, injecting both sketch semantic masks and keypoint heatmaps into each individual U-Net layer through the use of spatial structural constraints to ensure the generated image's structural accuracy.

User-input contour information is transformed through an alignment process achieved via pixel-level correspondence to create a control feature map  $F_c$ . The creative flow of ideas is shaped using textual clues and style memories allowing the intermediate layers' features to be modulated utilizing a cross-attention process to produce a feature map  $F_e$  containing artistic diversity. The style memory retrieval module extracts certain features from the artistic database associated with particular styles of art. A gated fusion network imposes dynamic balancing between both the control

and creative elements using a learnable spatial weight map  $W$  (ranging from 0 to 1) that is constructed through an individual element-by-element weighted fusion process. The weight map is produced through a convolutional-sigmoid architecture applying different fusion tactics for various regions of the input image. An iterative scheduler processes these weights during the control stages to determine the relative weight of control vs. creativity during different times of the experiment; A global temporal cosine annealing scheduler multiplies the local spatial weight map  $W$  at each denoising step to dynamically adjust the overall balance between divergent artistic exploration and convergent structural alignment.

TABLE 1. Core components of this framework and the technical solutions adopted

| Core Module                 | Technical Goal                                    | Input Data                                          | Key Technical Scheme                                                                       | Output Form          |
|-----------------------------|---------------------------------------------------|-----------------------------------------------------|--------------------------------------------------------------------------------------------|----------------------|
| Precise Control Stream      | Structural fidelity and semantic alignment        | Sketch semantic mask, keypoint heatmap, text vector | ControlNet-like conditional guidance with spatial structure injection at every U-Net layer | Control feature map  |
| Creative Stimulation Stream | Artistic expression and style diversity           | Text prompt, style-memory retrieval vector          | Cross-attention modulation of diffusion mid-layer features                                 | Creative feature map |
| Adaptive Fusion Gate        | Adaptive trade-off between control and creativity | $F_c, F_e$                                          | Concatenate $\rightarrow$ Conv $\rightarrow$ Sigmoid to generate spatial weight map        | Weight map           |
| Dynamic Fusion Strategy     | Region and stage-adaptive fusion                  | $W, F_c, F_e$                                       | Element-wise fusion                                                                        | Fused feature        |
| Iterative Weight Scheduler  | Step-wise creativity-control balance              | Current time step                                   | Linear/cosine/piece-wise increase of weight                                                | Spatial weight map   |

This section illustrates how the proposed fusion method enables a transition from disjointed creation to collaborative optimization. The approach resolves the tension between semantical accuracy and artistic creativity while establishing a human-machine collaboration framework for continuous

iterative artistic generation, thereby providing a practical and innovative technical pathway for producing professional-quality concept art.

### III. EXPERIMENT AND RESULT ANALYSIS

#### A. EXPERIMENTAL DATASET

This study was primarily conducted using data from Wikimedia Commons. All images have been screened through Wikimedia's advanced search functions to only include High-Resolution Concept Art that is either Public Domain or Creative Commons Licensed Images. The types of images included cover an array of topics such as character design, game scene design, and Science Fiction city designs. All images were carefully evaluated by hand to ensure the imagery being selected was appropriate for this study, as well as to verify the image participant had the appropriate rights to allow for that image to be used in this study.

Using the Python Scrapy Framework, we collected the images and associated metadata to build the data sets that will be used in this study. A team of Art Professionals conducted two rounds of review to ensure all reviewed images of each of the collected image groups meet the requirements of the study, including quality, copyright compliance, and content consistency. Using the original tag of the images, we have created manually structured image-text pairs for each of the categories included in our Data Set for the experiment detailed in this report; each category contains between 5,000 high-quality images, and approximately 1,600-1,700 images per category. All images have been provided with a text description (i.e., Female Warrior in Eastern Fantasy Style, Light Armor, Magical Sword in Floating Mountains) that helps explain the contents of the image as well as the style of art. 80% of the dataset is used for style memory training and model fine-tuning, while the remaining 20% is used as a test set for final evaluation, ensuring that the test samples are not involved in any training or encoding process.

Table 2 shows an overview of the dataset, which is divided into three major task categories: character design, scene composition, and prop setting. It clearly presents the data split results: the training set consists of 3,500 images (70%), the validation set 500 images (10%), and the test set 1,000 images (20%).

Among them, the character design sample size is the largest, followed by the scene and prop setting, and the overall distribution is balanced.

**TABLE 2.** Dataset overview

| Split      | Character Design | Scene Composition | Prop Design | Total |
|------------|------------------|-------------------|-------------|-------|
| Train      | 1190             | 1155              | 1155        | 3500  |
| Validation | 170              | 165               | 165         | 500   |
| Test       | 340              | 330               | 330         | 1000  |

#### B. EXPERIMENTAL SETUP

The comparative experiments select the following mainstream generative models as baselines:

(1) Stable Diffusion + ControlNet: Using the Canny control mode, with sketches and text as input, it represents a highly controllable generative method;

(2) Midjourney v6: Using the official API with default stylize parameters and identical aspect ratios under the same text prompt and sketch weight parameters;

(3) Text-to-Image with LoRA: Based on Stable Diffusion, it adds the LoRA module of a specific artist style to evaluate the style reproduction ability. All models are run on the same hardware environment, and the generated image resolution is uniformly 768×768.

In terms of input conditions, the test tasks are designed into three typical creative scenarios: character design (including pose constraints), scene atmosphere exploration, and cross-style transfer. Each task has the same prompt and includes text-based instructions, drawings made with a pen, and examples of other works with similar styles, so all the comparisons can be done fairly. In creating this framework, the style memory library was pre-loaded with 20 different styles of art (for example, Ink Painting Art and Low Polygon Art), and therefore the automatic retrieval/injection of styles allowed user-to-browser matching of art styles.

To evaluate the system, the evaluation is based on both objective and subjective measures, as shown in Table 3. The objective measures are: Frechet Inception Distance (FID), which quantifies the overall quality of the generated image and how close it is to the distribution of actual fine art; and CLIP Score, which quantifies how semantically consistent the generated image is with the text input's semantic meaning/content accuracy.

The subjective evaluation involves a double-blind test conducted by 150 professional concept artists with at least three years of experience.

**TABLE 3.** Scoring criteria for subjective evaluation dimensions

| Dimension                  | Definition                                                                                            | Scoring Guidelines                                                                                                                                                                                                                                                     |
|----------------------------|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Semantic Accuracy          | How well does the image match the input text description and sketch layout?                           | 5: all key objects and relationships are correctly depicted; 4: only imperceptible deviations; 3: main content is correct but minor details are missing or inaccurate; 2: key object missing or wrong; 1: major elements are missing or fundamentally incorrect.       |
| Structural Controllability | How precisely is the spatial structure preserved, such as proportions, pose, and component placement? | 5: body ratios and object arrangements strictly follow the sketch; 4: error 3%–8%, visible only on overlay; 3: 8%–15% drift, structure still recognizable; 2: 15%–25% distortion, limbs misplaced, perspective clashes; 1: >25% warping, broken skeleton, unmatchable. |
| Artistic Creativity        | How novel and imaginative is the visual expression?                                                   | 5: highly original and inspiring, beyond generic patterns; 4: 1–2 novel touches within mainstream frame; 3: some creative elements but largely conventional; 2: surface tweaks only; 1: boring, repetitive, or lacking artistic flair.                                 |
| Style Consistency          | Are the artistic style, brushwork, color palette, and lighting coherent across the entire image?      | 5: uniform and harmonious throughout; 4: micro-hop visible only zoomed-in; 3: mostly consistent with minor local inconsistencies; 2: two styles co-exist with harsh transition; 1: style conflicts or jarring transitions.                                             |
| Detail Richness            | How rich and meaningful are the non-core visual details?                                              | 5: abundant details enhancing narrative depth; 4: adequate detail with 1–2 minor empty patches; 3: adequate but unremarkable; 2: sparse or mechanical repetition; 1: overly simplistic, empty, or noisy.                                                               |
| Visual Harmony             | Is the overall composition balanced, aesthetically pleasing, and free of visual clashes?              | 5: professionally composed with excellent color balance; 4: globally harmonious with tiny local bump; 3: main content is correct but minor details are missing or inaccurate; 2: multiple conflicts; 1: major elements are missing or fundamentally incorrect.         |
| Overall Usability          | Can the image be immediately inserted into the downstream design pipeline without extra labor?        | 5: ready for final production without retouch; 4: usable after <1 min micro-edits; 3: usable with ≤10% rework; 2: requires >10% but <50% rework; 1: fundamental flaws, quicker to re-generate than to fix.                                                             |

Each reviewer scored each generated result on seven dimensions (1-5 points):

- (1) Semantic Accuracy (whether the content matches the input);
- (2) Structural Controllability (whether the sketch structure is faithfully restored);
- (3) Artistic Creativity (whether the visual expression is novel and unique);
- (4) Stylistic Consistency (whether the style of the picture is unified);
- (5) Richness of Detail (whether the texture and background are substantial);
- (6) Visual Harmony (whether the composition and color are balanced and beautiful);
- (7) Overall Usability (whether the results can be directly used in subsequent design processes), which systematically measures the capability of the generated results in terms of fidelity and creativity. Each dimension is accompanied by a clear operational definition and a 5-point scoring standard to ensure consistent understanding among different reviewers, thereby obtaining reliable and analyzable feedback data.

#### IV. RESULTS

This section presents examples of concept art output by the system under typical input conditions, and compares them with baseline models. All tests are based on the same input of text prompts, hand-drawn sketches, and style references to ensure consistency and fairness. Examples of multimodal input conditions for testing are shown in Figure 4.

Figure 4 is divided into three test cases. The first task is to create a “Future Mech Warrior” character, with input being a description (“A male mech warrior with back thrusters

and energy rifle who is wearing modular armor in front of a ruined city.”) plus a pose sketch and images for inspiration in Cyberpunk-style.

The second is about making an “Oriental Fantasy Forest Temple” based on the text from the task of “arranged stone pillars covered with mist”, which gives direction to the overall composition and design of the main scene. The third task is to create two different interpretations of one character in two different styles. One interpretation should be in a realistic oil painting style, and the second should be in Japanese Cel-Shaded Style for the same character and background.

The resulting concept art output from various mainstream models is shown in Figure 5.

As shown in Figure 5, the comparison of concept art results shows that in the first case study, the character generated by our proposed method has accurate structural proportions, and the armor details are highly consistent with the sketch layout. The metallic texture of the background, combined with the neon colors, give off a strong industrial/toughness and stylized expression, while enhancing the depth of the overall artwork. By contrast, while Stable Diffusion + ControlNet has structural stability, its color palette is very monotone, and the background is simplified, and therefore lacks an artistic aspect. Midjourney also had issues with simplifying the background, but in addition to that, it produced a weapon with an alternate shape from the intended sketch. Although the LoRA fine-tuned model has a similar style, it is off due to misalignments between components.

When comparing the results from the second set of illustrations to the original illustration, there are many similarities in the arrangement of the three stone pillars and the direction of the path taken. However, the style memory technique successfully created the artistic texture (ink wash) and light overlay color of the original illustrations, giving it both a sense of pre-planning and poetic representation. The baseline model primarily used the ControlNet model to recreate the watercolor effect and ignored the ink wash. Additionally, Midjourney experienced the same problems; buildings appeared to float illogically and had an inaccurate proportion to one another. Like Midjourney, the LoRA method was limited by the amount of training data available; this resulted in missing some detail in the original illustrations, making it hard to generalize on complex compositions.

The third set of this framework achieves high-quality style switching by calling the corresponding vectors in the style memory library. The main structure of the character remains stable, the posture remains unchanged, and the background remains basically unchanged except for the color difference, verifying the ability of style decoupling and flexible calling.

A comprehensive analysis and comparison of various mainstream generative models shows that although Stable Diffusion + ControlNet can ensure structural alignment through conditional injection, its single-stream architecture

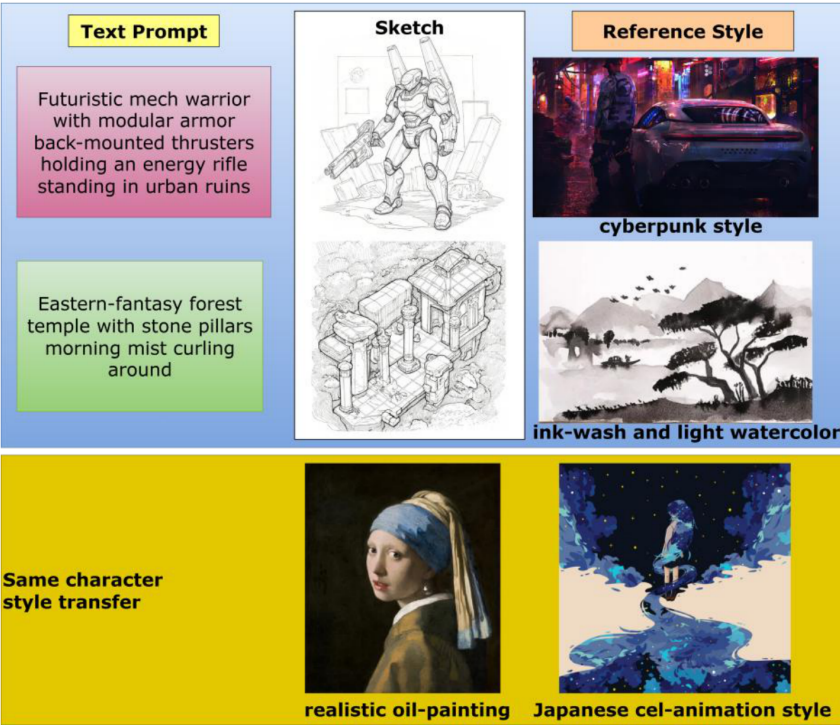


FIGURE 4. Example of multimodal input conditions (text, sketch, style reference image)

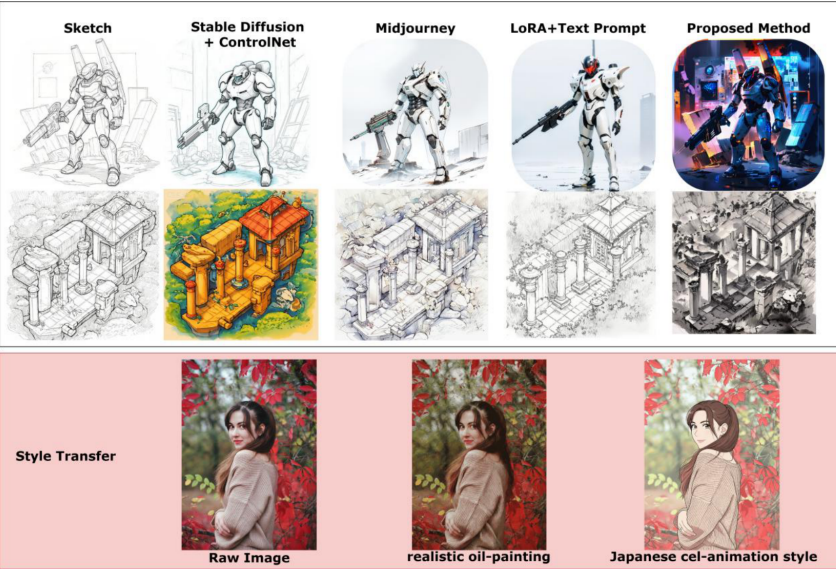


FIGURE 5. Comparison of art concept maps generated by different methods

fuses control signals with fixed weights [32] and lacks regional adaptive capabilities, resulting in over-emphasis on geometric consistency and suppression of visual diversity. As a result, the generated results are often dull and monotonous in color, and cannot actively stimulate stylized texture and atmosphere rendering. Although commercial models such as Midjourney have strong aesthetic priors, they rely on text and low-resolution sketch inputs, making it difficult to accurately analyze the fine-grained layout and spatial relationships in hand-drawn drafts. They lack explicit

structural supervision [33] and are prone to problems such as floating buildings, unbalanced proportions, and misplaced components. This reflects the model's preference for global aesthetics first rather than local fidelity first and explains why it has been criticized for its limited ability to control fine details [34]. The LoRA fine-tuning method is limited by the distribution of training data [35]. It is essentially a low-rank correction of the parameter space and has weak generalization ability. When faced with new compositions or cross-pose tasks, details are often missed or styles are

confused. In addition, the content and style are tightly coupled, and there is a lack of decoupling mechanism, making it difficult to achieve flexible style switching. These methods generally have three major problems: coarse control granularity, rigid style expression and weak multimodal understanding. They are unable to achieve fine-grained control over shape preservation in key areas and artistic generation in non-key areas, lack support for searchable and editable external style memory, and fail to establish a strong connection between sketch structure and text semantics.

The proposed method achieves a good balance between structural fidelity and artistic expression, significantly outperforming baseline models in terms of edge detail processing, light and shadow layering, and stylistic consistency. Designers can also adjust fusion weights or replace style references to obtain new variants in real time, significantly improving creative efficiency and the breadth of exploration.

### A. ABLATION EXPERIMENT

This section evaluates the impact of multimodal input conditions, style memory, and dynamic fusion strategies on semantic accuracy and artistic creativity. All experiments are conducted on the same test set, using FID, CLIP Score, and subjective ratings as evaluation metrics.

The baseline model is Stable Diffusion (v2.1), driven solely by textual cues and without any structural or stylistic control. Building upon this, modules are gradually added:

Variant A: Baseline + Sketch Structure Constraints (Precise Control Flow): Incorporates sketch information from semantic segmentation and pose extraction into the generation process via a ControlNet mechanism.

Variant B: Baseline + Style Memory (Creative Inspiration Flow): Builds upon the baseline model with a style encoder and retrieval mechanism, using the matching style vectors to guide generation.

Variant C: Baseline + Dynamic Fusion Mechanism (Full Framework): Enables both the Precise Control Flow and the Creative Inspiration Flow, and achieves adaptive fusion via a gating network.

TABLE 4. Ablation experiments

| Model Configuration | Module Combination                                   | FID   | CLIP Score |
|---------------------|------------------------------------------------------|-------|------------|
| Baseline Model      | Text prompt Only                                     | 35.60 | 0.278      |
| Variant A           | Sketch structure constraint (precise control stream) | 24.80 | 0.293      |
| Variant B           | + Style memory bank (creative inspiration stream)    | 37.20 | 0.281      |
| Variant C           | Dual-stream fusion                                   | 18.73 | 0.321      |

According to the ablation results in Table 4, the baseline model, relying solely on textual cues, achieved an FID of 35.60 and a CLIP score of 0.278. When sketch or pose structures are not clearly defined, the generated images demonstrate some deviance from original intention and lack effective visual representation. The inclusion of both

structure and pose in Variant A provides a high degree of semantic consistency (measured by lower FID score) of 24.80, and an increase of CLIP score (based on human evaluation), at 0.293.

When sketch/pose and style memory paths are maintained as in Variant B, there is no longer an accurate representative path resulting in a slight increase in FID score (37.20) among the generated images. The isolated injection of style memory without structural constraints disrupts the original spatial coherence of the baseline model and causes a deterioration in the Fréchet Inception Distance metric. In the full framework (Variant C), both paths are utilized through a dynamic fusion mechanism, achieving the highest objective performance through a lower FID of 18.73 and a CLIP score (from human evaluation) of 0.321.

In summary, the precise control flow maximizes semantic accuracy, while the style memory enhances the visual expressiveness and creativity. Combining the two will allow for the merging of these creative signals through a dual pathway system, without compromising either. The collaboration between the three elements of this framework will allow for a highly flexible process to generate quality artwork concept art, which satisfies the two primary requirements of professional artists.

### B. RESULTS ANALYSIS

By comprehensively analyzing the objective indicators and comparing the FID and CLIP Score of each mainstream model, the evaluation indicators of each mainstream model are shown in Figure 6.

In terms of objective quantitative metrics, this framework achieves an FID of 18.73 on the test set, significantly outperforming baseline models (Stable Diffusion + ControlNet: 24.80; Midjourney: 20.10; LoRA:

23.50), demonstrating that its generated images are superior in terms of visual quality and closeness to the distribution of real-world art. Furthermore, its CLIP score reaches 0.321, exceeding all other compared models, demonstrating more accurate semantic alignment between text and image, and a more faithful representation of design intent. This result validates the effectiveness of multimodal conditional input and the dual-pathway architecture in information fusion.

Figure 7 shows a radar chart of subjective user feedback from a user study. The chart utilized the multi-dimensional subjective evaluation results from 150 professional concept artists regarding the proposed generation framework (the two models being evaluated) against the baseline model (Stable Diffusion + ControlNet). The evaluations looked at seven primary dimensional categories: 1) Semantic Accuracy, 2) Structural Controllability, 3) Artistic Creativity, 4) Stylistic Consistency, 5) Richness of Detail, 6) Visual Harmony, and 7) Overall Usability, where Higher Scores indicate a more favorable response of an artist.

The framework received a rating of 4.4 for the best Structural Controllability, meaning the framework is capa-

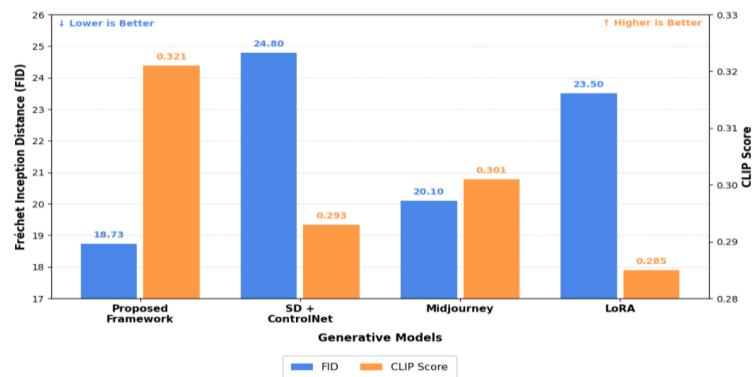


FIGURE 6. Evaluation comparison of various mainstream model methods

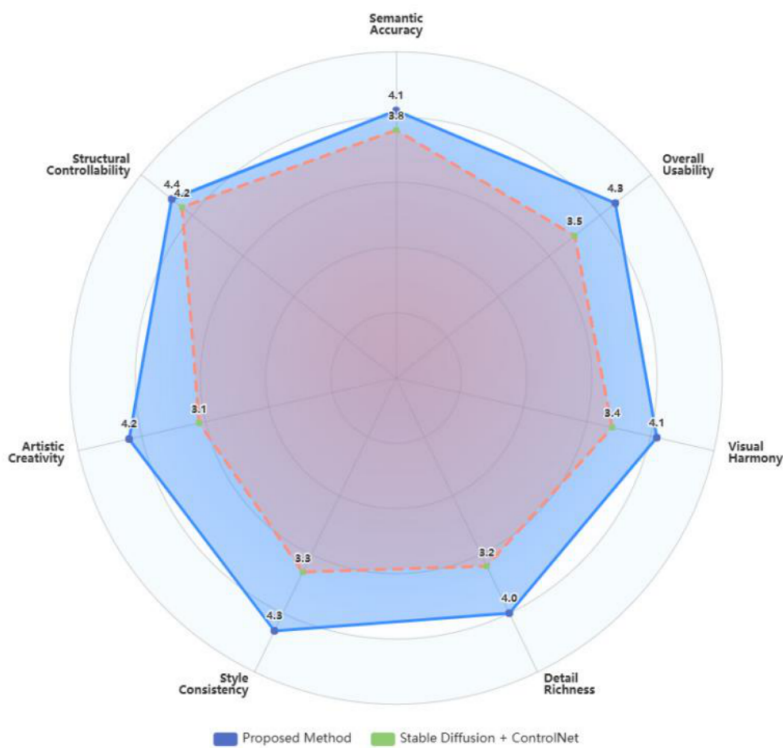


FIGURE 7. User research feedback radar chart

ble of locking onto the structure of the artists’ sketches; therefore, the Framework Improved upon the baseline model substantially regarding Artistic Creativity (4.2) and Stylistic Consistency (4.3); therefore, the results lend support to the Style Memory’s effectiveness. Additionally, the Baseline Model performed well in the Structural Control dimension but received lower ratings than the proposed framework in the Artistic Expression Dimensions (Artistic Creativity and Stylistic Consistency), indicating that the Baseline Model produced artwork that is more mechanical and less aesthetically pleasing.

In conclusion, from the experimental results, we find that the Framework increased the artistry and usability of the

generated content while providing the artist with a high level of control through multimodal conditional driving, style memory guidance and dynamic fusion regulation. The result is an improved, reliable, and creative technical method for developing intelligent art.

V. CONCLUSION

This study proposes a dual-path collaborative generation framework for AI-driven concept design using multimodal inputs, including textual prompts and hand-drawn sketches. This is done using a combination of precision control flow (the structured way the program processes drawings) and the creative inspiration flow (the programming that

generates the creative design content based on previously examined styles). A gating network is used to dynamically combine these two production flows, using weights that can be dynamically adjusted. Based on this framework's performance in these experiments, it outperformed all other baseline models currently being used in AI-driven concept generation using objective measures (e.g.

FID (18.7) and CLIP Score (0.321)). Additionally, the accuracy, creativity, and usability of all generated designs using this framework were rated significantly higher compared to previous models by professional concept artists based on subjective evaluations. Given the high performance in balancing structural constraints with artistic style, the model demonstrates strong applicability to industries requiring precise layout control applied to digital textile pattern design and fabric visualization. Results from ablation studies support the successful interaction and combination of both modules (gating network and style memory) and how these systems produce the best quality images. This integrated paradigm drives the digital transformation of the textile industry by maintaining technical fabric structures alongside the exploration of diverse creative motifs. More broadly speaking, this paper details a new method by which professionals can fully leverage generative AI in professional arts, opening the door to the next evolution of human and machine collaboration from a tooling approach (e.g., artists using technology as tools) to an integrated approach (e.g. artists creating together with technologists).

### AUTHOR CONTRIBUTIONS

Miao Hu is responsible for all the work of the article.

### CONFLICTS OF INTEREST

The author declares that there are no conflicts of interest regarding the publication of this article.

### FUNDING

The author received no financial support for the research, authorship, and/or publication of this article.

### ACKNOWLEDGEMENTS

Not applicable.

### REFERENCES

- [1] M. Engler and A. Trnka, "Fundamental Knowledge behind Creation of Concept Art," *Eur. J. Media, Art Photogr.*, vol. 9, no. 2, pp. 106-115, 2021.
- [2] Z. Epstein, A. Hertzmann, L. Herman, R. Mahari, R. Frank, M. Groh, et al., "Art and the science of generative AI," *Science*, vol. 380, no. 6650, pp. 1110-1111, 2023, doi: 10.1126/science.adh4451.
- [3] I. O. Adeleye, "The impact of artificial intelligence on design: Enhancing creativity and efficiency," *Journal of Engineering and Applied Sciences*, vol. 3, no. 1, pp. 1-13, 2024, doi: 10.70560/vvsfej12.
- [4] Y. Wei and Z. Zhao, "Integration effect of artificial intelligence and traditional animation creation technology," *Journal of Intelligent Systems*, vol. 33, no. 1, pp. 20230305-20230305, 2024, doi: 10.1515/jisys-2023-0305.
- [5] T. Chandrasekera, Z. Hosseini, and U. Perera, "Can artificial intelligence support creativity in early design processes?," *International journal of architectural computing*, vol. 23, no. 1, pp. 122-136, 2025, doi: 10.1177/14780771241254637.
- [6] K. O'Toole, E. A. Horvat, "Extending human creativity with AI," *Journal of Creativity*, vol. 34, no. 2, pp. 100080-100080, 2024, doi: 10.1016/j.jyoc.2024.100080.
- [7] 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-e16757, 2023, doi: 10.1016/j.heliyon.2023.e16757.
- [8] V. Paananen, J. Oppenlaender, and A. Visuri, "Using text-to-image generation for architectural design ideation," *International Journal of Architectural Computing*, vol. 22, no. 3, pp. 458-474, 2024, doi: 10.48550/arXiv.2304.10182.
- [9] S. Lee, M. Law, and G. Hoffman, "When and how to use AI in the design process? Implications for human-AI design collaboration," *International Journal of Human-Computer Interaction*, vol. 41, no. 2, pp. 1569-1584, 2025, doi: 10.1080/10447318.2024.2353451.
- [10] Y. Jin, J. Yoon, J. Self, and K. Lee, "Understanding Fashion Designers' Behavior Using Generative AI for Early-Stage Concept Ideation and Revision," *Archives of Design Research*, vol. 37, no. 3, pp. 25-45, 2024, doi: 10.15187/adr.2024.07.37.3.25.
- [11] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, pp. 100004-100004, 2021, doi: 10.1016/j.jjimei.2020.100004.
- [12] H. W. L. Mak, R. Han, and H. H. F. Yin, "Application of variational autoEncoder (VAE) model and image processing approaches in game design," *Sensors*, vol. 23, no. 7, pp. 3457-3457, 2023, doi: 10.3390/s23073457.
- [13] B. Y. Archana, "An analysis on the use of image design with generative AI technologies," *International Journal of Trend in Scientific Research and Development*, vol. 8, no. 1, pp. 596-599, 2024.
- [14] H. Li, L. Wang, and J. Liu, "A review of deep learning-based image style transfer research," *The Imaging Science Journal*, vol. 73, no. 4, pp. 504-526, 2025, doi: 10.1080/13682199.2024.2418216.
- [15] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, et al., "Dall-e: Creating images from text," *UGC Care Group I Journal*, vol. 8, no. 14, pp. 71-75, 2021.
- [16] L. Tan and M. Luhrs, "Using Generative AI Midjourney to enhance divergent and convergent thinking in an architect's creative design process," *The Design Journal*, vol. 27, no. 4, pp. 677-699, 2024, doi: 10.1080/14606925.2024.2353479.
- [17] C. Zou and Y. Chen, "A Study of Controllability Paths for Generative AI Design Based on Stable Diffusion ControlNet," *International Journal for Housing Science and Its Applications*, vol. 46, no. 3, pp. 2321-2332, 2025, doi: 10.70517/ijhsa463194.
- [18] Y. Chen and H. Ruan, "Deep Analogical Generative Design and Evaluation: Integration of Stable Diffusion and LoRA," *Journal of Mechanical Design*, vol. 147, no. 5, pp. 051403-051403, 2025, doi: 10.1115/1.4066861.
- [19] D. Choi, J. Im, and Y. Sung, "LoRA Fusion: Enhancing Image Generation," *Mathematics*, vol. 12, no. 22, pp. 3474-3474, 2024, doi: 10.3390/math12223474.
- [20] C. Shen, S. Luo, L. Fan, and C. Dai, "Leveraging AI and diffusion models for anime art creation: A study on style transfer and image quality evaluation," *Computer Science and Information Systems*, vol. 22, no. 3, pp. 1331-1346, 2025, doi: 10.2298/csis241212045s.
- [21] G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, et al., "Semantic communications for artificial intelligence generated content (AIGC) toward effective content creation," *IEEE Network*, vol. 38, no. 5, pp. 295-303, 2024, doi: 10.48550/arXiv.2308.04942.
- [22] Y. Lou, "Human creativity in the AIGC era," *She Ji: The Journal of Design, Economics, and Innovation*, vol. 9, no. 4, pp. 541-552, 2023, doi: 10.1016/j.sheji.2024.02.002.
- [23] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1-14, 2024, doi: 10.1057/s41599-024-03811-x.
- [24] C. Jiang, R. Huang, and T. Shen, "Generative AI-Enabled Conceptualization: Charting ChatGPT's Impacts on Sustainable Service Design Thinking With Network-Based Cognitive Maps," *Journal of Computing and Information Science in Engineering*, vol. 25, no. 2, pp. 021006-021006, 2025, doi: 10.1115/1.4067211.
- [25] H. U. Zi-Yang, "AIGC related context: A new communication culture for human," *Journal of Literature and Art Studies*, vol. 14, no. 10, pp. 921-931, 2024, doi: 10.17265/2159-5836/2024.10.016.

- [26] L. Abrusci, K. Dabaghi, S. D'Urso, and F. Sciarrone, "AI4Design: A generative AI-based system to improve creativity in design—A field evaluation," *Computers and Education: Artificial Intelligence*, vol. 8, no. 1, pp. 100401-100401, 2025, doi: 10.1016/j.caeai.2025.100401.
- [27] F. Jiang, C. Tang, L. Dong, K. Wang, K. Yang, and C. Pan, "Visual Language Model-Based Cross-Modal Semantic Communication Systems," *IEEE Transactions on Wireless Communications*, vol. 24, no. 5, pp. 3937-3948, 2025, doi: 10.1109/TWC.2025.3539526.
- [28] Y. Alaluf, E. Richardson, G. Metzger, and D. Cohen-Or, "A neural space-time representation for text-to-image personalization," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1-10, 2023, doi: 10.48550/arXiv.2305.15391.
- [29] N. Manakitsa, G. S. Maraslidis, L. Moysis, Fragulis, and GF, "A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision," *Technologies*, vol. 12, no. 2, pp. 15, 2024, doi: 10.3390/technologies12020015.
- [30] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Explicit Filterbank Learning for Neural Image Style Transfer and Image Processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2373-2387, 2021, doi: 10.1109/tpami.2020.2964205.
- [31] K. Hu, Z. Wang, K. A. E. Martens, M. Hagenbuchner, M. Benmamoun, and A. C. Tsoi, "Graph fusion network-based multimodal learning for freezing of gait detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1588-1600, 2021, doi: 10.1109/TNNLS.2021.3105602.
- [32] L. G. Foo, H. Rahmani, and J. Liu, "Ai-generated content (aige) for various data modalities: A survey," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1-66, 2025, doi: 10.48550/arXiv.2308.14177.
- [33] A. W. Sukkar, M. W. Fareed, M. W. Yahia, S. B. Abdalla, I. Ibrahim, and K. A. K. Senjab, "Analytical evaluation of Midjourney architectural virtual lab: Defining major current limits in AI-generated representations of Islamic architectural heritage," *Buildings*, vol. 14, no. 3, pp. 786-786, 2024, doi: 10.3390/buildings14030786.
- [34] G. Buzzaccarini, R. S. Degliuomini, M. Borin, A. Fidanza, N. Salmeri, L. Schiraldi, et al., "The promise and pitfalls of AI-generated anatomical images: evaluating midjourney for aesthetic surgery applications," *Aesthetic plastic surgery*, vol. 48, no. 9, pp. 1874-1883, 2024, doi: 10.1007/s00266-023-03826-w.
- [35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, pp. 3-3, 2022, doi: 10.48550/arXiv.2106.09685.