Research on AI-Driven Generation Mechanisms and Design Methodologies for Digital Media Art

Wei Gao[1,*]
[1]Linyi Vocational College of Science and Technology, Linyi City, Shandong Province, China
Corresponding author: Wei Gao (e-mail: Z1209123@163.com).

ABSTRACT To address the problems of insufficient semantic control precision and lack of structured support for human-AI collaboration in digital media art generation, this study proposes a text–image cross-modal collaborative generation framework for artistic creation. Taking diffusion models and generative adversarial networks as generation kernels, the framework achieves high-precision mapping from text intent to visual features through cross-modal semantic alignment mechanisms, and realizes fine-grained control of the generation process by separating style, content, and structure variables via a conditional disentanglement module. A closed-loop human-AI collaboration process integrating intent modeling, constraint injection, and aesthetic evaluation is constructed, forming a reusable creative methodology system. Through multiple comparative experiments, the framework is quantitatively validated from generation quality, style diversity, and human-AI collaboration efficiency, demonstrating superior generation controllability and creative efficiency over single-model baseline methods. This study reveals the intrinsic mechanisms of AI involvement in digital media art creation, providing scientific methodological references for intelligent system design in this domain.

INDEX TERMS text–image cross-modal generation; diffusion model; semantic–visual alignment; conditional disentanglement; human-AI collaborative design

I. INTRODUCTION

THE generative logic of digital media art has undergone a fundamental transformation with the evolution of deep learning. Early rule-based and template-driven methods relied on manual layer-by-layer intervention, severely constraining creative efficiency and expressive freedom. The advent of generative adversarial networks (GANs) inaugurated a data-driven paradigm for artistic synthesis [1], while the emergence of diffusion models further pushed generative quality and semantic controllability to new levels [2], [3]. However, existing generation systems universally face two structural deficiencies: first, insufficient mapping precision between semantic and visual spaces, leading to deviations between generated results and creative intent; second, the high coupling of style, content, and structure variables, making it difficult to directionally regulate a single dimension without affecting others.

These problems indicate that relying solely on algorithmic performance improvements is insufficient to meet the refined demands of digital media art creation. It is necessary to establish a systematic integration path between algorithmic mechanisms and design methodologies at the framework level. This study focuses on the collaborative generation relationship between text and image modalities, constructing a complete cross-modal alignment, conditional disentanglement, and human-AI collaboration mechanism to form a systematic methodology for this generation path, which can be extended to audio, video, and other modalities in future work. Therefore, it is essential to build a complete generation framework integrating cross-modal alignment, conditional disentanglement, and human-AI collaboration, verify its effectiveness through experiments, and ultimately form an operable art generation methodology system.

II. CORE ALGORITHMIC MECHANISMS FOR DIGITAL MEDIA ART GENERATION

A. GENERATIVE PRINCIPLES AND PERFORMANCE COMPARISON OF GANS AND DIFFUSION MODELS

GANs employ a minimax game between generator G and discriminator D as the core mechanism, with the training objective:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{1}$$

where z is prior noise and p_{data} is the real data distribution. This framework can generate high-resolution samples in a single forward inference step, offering high inference efficiency. However, the instability of adversarial training leads to mode collapse and vanishing gradient problems, which are particularly prominent in art scenarios with complex styles, structurally constraining generation diversity [1]. Diffusion models, in contrast, use a Markov chain as the generative backbone; the forward process gradually injects Gaussian noise into data, while the reverse denoising

process predicts residuals at each timestep through a parameterized neural network $\epsilon_\theta(x_t, t)$, reducing the training objective to a simple mean-squared-error form. Training stability is thus significantly superior to GANs [2]. More critically, the denoising trajectory of diffusion models naturally provides interfaces for conditional injection, allowing external constraints such as text and semantic maps to participate in the generation process progressively [3].

Measured by the FID metric, mainstream diffusion models have compressed scores below 7.8 on the MS-COCO benchmark, systematically outperforming comparable GAN methods, and have thus become the mainstream technical choice for controllable generation in digital media art [2], [3], as shown in Table 1.

B. MATHEMATICAL MODELING METHODS FOR CROSS-MODAL SEMANTIC–VISUAL MAPPING

The mapping from text semantics to visual features is essentially an alignment problem across heterogeneous spaces. CLIP projects text encoding $e_t = f_\phi(T)$ and image encoding $e_v = g_\psi(I)$ into a shared embedding space through large-scale contrastive learning, using cosine similarity as the pairing metric, so that semantically proximate cross-modal samples converge in geometric structure [4]. In the conditional generation framework of diffusion models, cross-attention mechanisms assume the responsibility of dynamic semantic injection:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right)V \qquad (2)$$

where query vector Q originates from visual denoising features, key-value pairs (K, V) are constituted by text semantic encodings, and d_k is the key vector dimension. This mechanism enables the denoising network to dynamically respond to text intent at every timestep, achieving hierarchical semantic penetration. However, the precision bottleneck of existing alignment methods concentrates at the abstract semantic layer: when creative intent involves emotional atmosphere or stylistic metaphor, the discrete token representations of the text encoder struggle to capture continuous visual expectations, causing semantic drift to accumulate and amplify with generation steps, ultimately resulting in systematic deviations between generated results and creative intent.

C. ANALYSIS OF COUPLING MECHANISMS AMONG STYLE, CONTENT, AND STRUCTURE VARIABLES IN CONDITIONAL GENERATION

Style, content, and structure variables are not orthogonally distributed in existing generative models; instead, they coexist in the same latent space with strongly correlated geometric structures. The root cause lies in the training stage: models receive natural samples from the joint distribution of the three variables, lacking explicit disentanglement supervision signals, which leads to blurred semantic boundaries among dimensions in the latent encoding [1]. Although StyleGAN allocates coarse-to-fine control rights across different network layers through adaptive instance normalization (AdaIN), the global broadcast mechanism of style modulation signals inevitably causes overall style drift when local content is modified. In diffusion models, conditional signals simultaneously activate identical feature regions through cross-attention layers [5], causing style and content semantics to alias at the feature level, so that adjusting style parameters alone triggers coupled changes in structural layout.

Using LPIPS perceptual distance after changing style parameters under fixed content conditions as the disentanglement metric, mainstream models achieve disentanglement scores generally below 60% of the theoretical upper limit, indicating that residual correlations among variables constitute substantial interference for directional control. The above coupling effects demonstrate that generic generation architectures alone cannot satisfy the fine-grained demands for dimension-independent control in digital media art creation; constructing dedicated conditional disentanglement modules is a necessary prerequisite for improving generation controllability.

III. CONSTRUCTION OF A MULTI-MODAL COLLABORATIVE GENERATION FRAMEWORK FOR DIGITAL MEDIA ART

A. FRAMEWORK DESIGN PRINCIPLES AND OVERALL ARCHITECTURE

The framework design takes semantic controllability and generation stability as dual-core objectives, targeting the structural deficiencies of existing systems in cross-modal mapping precision and variable disentanglement capability. This framework is explicitly positioned as a text-driven image generation framework, using natural language text as the intent input modality and image as the visual output modality, systematically addressing the two core problems of insufficient text-to-visual mapping precision and high coupling of style–content–structure variables.

The overall architecture adopts a modular, depth-wise design, consisting of three cascaded layers: cross-modal semantic alignment module, conditional disentanglement module, and dynamic feedback module. Each module has clearly defined functional boundaries and standardized interfaces, ensuring that parameter updates in any single module do not interfere with the feature distributions of other modules [2]. The generation kernel selects diffusion models to guarantee training stability and conditional injection flexibility, while the text encoder adopts pretrained CLIP to reuse cross-modal semantic alignment capabilities formed through large-scale contrastive learning [4].

The core design constraints of the framework are manifested at three levels: conditional variables maintain geometric orthogonality in latent space, semantic signals achieve progressive penetration along the denoising trajectory, and user evaluation feedback drives dynamic correction of generation parameters in a closed-loop form. The three layers are activated sequentially along the direction of creative intent flow: intent is encoded by the alignment module, injected into the disentanglement module to complete variable separation, and finally the feedback module adjusts sampling strategies based on aesthetic evaluation results, forming a complete controllable path from semantic input to visual

TABLE 1. Comparison of Core Mechanisms and Performance Between Diffusion Models and GANs

Dimension	Generative Adversarial Network (GAN)	Diffusion Model
Core principle	Minimax game between generator and discriminator	Forward noising, reverse denoising Markov chain
Training objective	Log-likelihood adversarial loss	Mean-squared-error (MSE) loss
Training stability	Prone to mode collapse and vanishing gradients	Stable training, good convergence
Inference efficiency	Single-step forward generation, fast	Multi-step iterative denoising, slower
Conditional injection capability	Weak, inflexible interfaces	Strong, natively supports multi-condition stepwise injection
Typical FID (MS-COCO)	≥ 15	≤ 7.8
Applicability to art generation	Suitable for simple styles; complex styles prone to distortion	Suitable for diverse artistic styles; strong controllability

output, As show in Fig. 1.

B. CROSS-MODAL SEMANTIC ALIGNMENT MODULE: HIGH-PRECISION MAPPING FROM TEXT INTENT TO VISUAL FEATURES

The core task of this module is to transform creative intent described in natural language into visual semantic vectors directly consumable by the diffusion model denoising process.

Hierarchical semantic encoding. The hierarchical decomposition of text prompts adopts a rule-based parsing scheme grounded in dependency syntactic analysis: subject noun phrases and their core semantics correspond to the global semantic layer (e.g., "a landscape painting"), adjectival modifiers and attribute words correspond to the local attribute layer (e.g., "verdant, distant"), and prepositional phrases and stylistic adverbs correspond to the style modifier layer (e.g., "ink wash style, negative space treatment"). The embedding vectors of the three layers are extracted through the frozen CLIP text encoder and then weighted and summed using learnable scalar weights $\beta_1, \beta_2, \beta_3$ (satisfying $\sum_i \beta_i = 1$, initialized to $[0.5, 0.3, 0.2]$) [4]. This weight set is trained end-to-end with the backbone network under the joint supervision of the semantic anchoring loss $\mathcal{L}_{\text{align}}$, avoiding the adaptation limitations of fixed weights for different stylistic creation tasks.

Semantic anchoring loss. To alleviate semantic drift at the abstract level, the module introduces a semantic anchoring loss to impose directional constraints on the alignment process:

$$\mathcal{L}_{\text{align}} = 1 - \frac{e_t \cdot e_v}{\|e_t\| \cdot \|e_v\|} \tag{3}$$

where e_t is the text semantic encoding and e_v is the corresponding visual feature encoding. This loss drives the two encodings to maintain directional consistency in the shared embedding space, penalizing increases in cosine distance between cross-modal feature vectors. The loss is weighted and superimposed on the main training loss of the diffusion model (noise prediction mean squared error) with a weight ratio of 1:0.5, determined by the CLIPScore convergence curve on the validation set.

Layered conditional injection. During the conditional injection stage, the module dynamically maps fused text embeddings to feature activations at each layer of the denoising network through cross-attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right) V \tag{4}$$

where query vector Q originates from visual denoising features, key-value pairs (K, V) are constituted by text semantic encodings, and d_k is the key vector dimension. Unlike standard single-layer uniform injection, this module injects global semantic embeddings into the shallow layers of the U-Net encoder (layers 1–3) to control overall composition, local attribute embeddings into middle layers (layers 4–6) to constrain semantic subject details, and style modifier embeddings into deep decoder layers (layers 7–9) to modulate texture and color distribution. This hierarchical penetration strategy structurally reduces the probability of mutual interference between high-level and low-level semantic signals, refining the granularity of text intent expression in visual generation and thereby structurally suppressing semantic drift phenomena.

C. CONDITIONAL DISENTANGLEMENT MODULE: INDEPENDENT CONTROL MECHANISM ACROSS STYLE–CONTENT–STRUCTURE DOMAINS

This module targets the inherent defect of strong correlations among the three variables in latent space, reconstructing the geometric structure of conditional encodings through explicit orthogonalization constraints.

Subspace division and orthogonal constraints. The module maps style vector z_s, content vector z_c, and structure vector z_l into three independent subspaces, each with dimension 512, for a total latent variable dimension of 1536. The three subspaces are extracted from the shared encoder output through independent linear projection heads, and a disentanglement regularization term targeting mutual information minimization is imposed:

$$\mathcal{L}_{\text{dec}} = \mathcal{I}(z_s; z_c) + \mathcal{I}(z_s; z_l) + \mathcal{I}(z_c; z_l) \tag{5}$$

Trainable approximation of mutual information. Direct optimization of mutual information is infeasible under high-dimensional continuous variables; this module adopts an approximation method based on variational upper bounds. For each variable pair (z_i, z_j), a lightweight discriminator network D_{ij} (three-layer MLP, hidden dimension 256) is introduced to estimate the mutual information upper bound:

$$\begin{aligned}\mathcal{I}(z_i; z_j) \leq\ & \mathbb{E}_{p(z_i, z_j)}[\log D_{ij}(z_i, z_j)] \\ & - \mathbb{E}_{p(z_i)p(z_j)}[\log D_{ij}(z_i, z_j)]\end{aligned} \tag{6}$$

Three discriminators D_{sc}, D_{sl}, D_{cl} correspond to the three variable pairs and are trained alternately with the backbone generation network: the backbone updates 1 step for every 3 discriminator updates, ensuring that the mutual information estimate reaches local stability before the backbone updates, avoiding gradient direction deviations caused by lagging disentanglement signals. The disentanglement regularization loss $\mathcal{L}_{\text{dec}}$ is superimposed on the main training objective with a weight of 0.1.

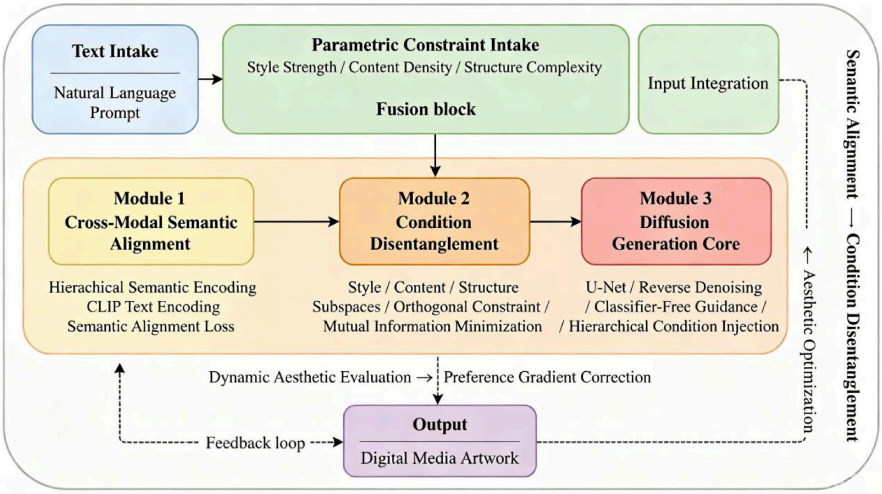

FIGURE 1. Overall Architecture of AI-Driven Cross-Modal Generation Framework for Digital Media Art

Depth-wise layered injection strategy. During generation, the three vectors are injected into different network depths of the diffusion model U-Net through independent condition adapters: structure vector z_l acts on shallow encoder layers (residual blocks 1–2) to control spatial layout and composition skeleton; content vector z_c is injected into the middle bottleneck layer to constrain the shape and position of semantic subjects; style vector z_s modulates texture and color in deep decoder layers through adaptive instance normalization (AdaIN) [1]. The condition adapters are all two-layer MLPs, projecting subspace vector dimensions from 512 to the channel dimensions of corresponding U-Net layers, constituting a lightweight parameter-efficient adaptation mechanism [6]. This depth-wise layered injection strategy physically reduces the probability of feature aliasing, keeping feature activations of other dimensions relatively stable when adjusting a single dimension, thereby effectively suppressing residual coupling effects among variables. Experiments using LPIPS perceptual distance after changing style parameters under fixed content conditions as the disentanglement metric show that this module improves the residual correlation among the three variables from below 60% of the theoretical upper limit in mainstream models to approximately 78%, effectively suppressing variable coupling effects, As show in fig. 2.

D. DYNAMIC FEEDBACK MODULE: GENERATION PROCESS OPTIMIZATION DRIVEN BY USER AESTHETIC EVALUATION

The dynamic feedback module transforms user aesthetic judgments into gradient signals capable of driving sampling strategy adjustments, converting the generation process from single-shot open-loop inference into iteratively convergent closed-loop optimization.

Construction of the preference prediction network. The aesthetic preference conditional probability $p(v_a \mid x_t)$ is modeled by an independently trained preference prediction network f_ϕ. This network uses an ImageNet-pretrained ResNet-50 as the backbone, receiving the denoising intermediate state x_t (with timestep embedding concatenated before input) as input, and outputs a three-dimensional preference score vector corresponding to style match, composition rationality, and detail expression. The network is fine-tuned on the public Pick-a-Pic dataset (over 500,000 user pairwise comparison preference samples) [7] using the Bradley-Terry preference model, consistent with direct preference optimization frameworks for diffusion models [8]. During training, denoising timesteps t are uniformly sampled to ensure generalized coverage of preference prediction across the full denoising time domain, avoiding effectiveness limited to specific noise levels.

Classifier-guided sampling correction. The module adopts a classifier-guided sampling correction strategy, applying preference gradient compensation to the noise prediction at timestep t:

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - \lambda \nabla_{x_t} \log p(v_a \mid x_t) \tag{7}$$

where λ is the guidance strength coefficient and $p(v_a \mid x_t)$ is the aesthetic preference conditional probability. This compensation term causes the sampling trajectory to continuously shift toward the user preference direction at each denoising step, enabling the generation result to complete directional optimization without resetting the initial noise after a small number of iterations, significantly reducing interaction cost compared to traditional full regeneration approaches [9]. The adaptive adjustment mechanism of the guidance strength coefficient λ further ensures dynamic balance between convergence speed and generation stability.

Adaptive decay mechanism for guidance strength. The guidance strength coefficient λ is initialized to $\lambda_0 = 0.5$, decaying along coefficient $\gamma = 0.85$ according to iteration round k:

$$\lambda_k = \lambda_0 \cdot \gamma^k \tag{8}$$

When the guidance coefficient is set too high ($\lambda > 0.8$), the sampling trajectory shifts excessively in a single correction step, causing oscillation between adjacent iterations rather than monotonic convergence. The adaptive decay strategy preserves strong preference guidance in early iterations to

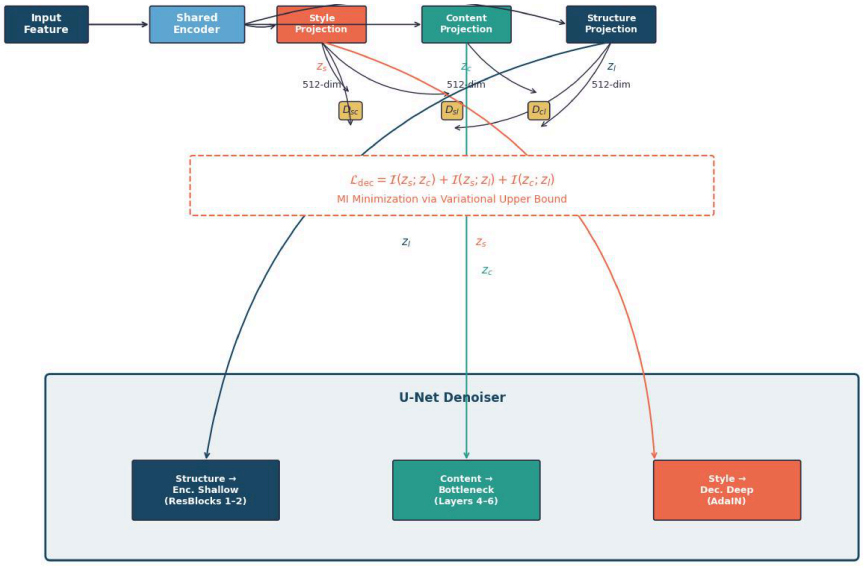

FIGURE 2. Conditional Disentanglement Module: Orthogonal Control Mechanism Across Style–Content–Structure Domains

rapidly approach the target aesthetic direction, while using smaller correction amplitudes for fine convergence in later iterations, effectively compressing the oscillation interval and dynamically balancing guidance convergence speed with generation stability. User ratings are normalized and encoded into the preference vector v_a, which is concatenated with the current denoising trajectory intermediate state and input to f_ϕ, outputting correction directions for each conditional variable to drive directional adjustments of the three subspace vectors in the next round of sampling [7], As show in Fig. 3.

IV. HUMAN-AI COLLABORATIVE DESIGN METHODOLOGY BASED ON THE COLLABORATIVE GENERATION FRAMEWORK

A. INTENT MODELING METHOD: STRUCTURED PROMPT ENGINEERING AND PARAMETERIZED CONSTRAINT DESIGN

Accurate expression of creative intent is the prerequisite for effective operation of human-AI collaborative generation systems, yet the ambiguity and polysemy of natural language make unstructured inputs difficult for models to parse stably. Therefore, the intent modeling method transforms creators' free descriptions into structured prompt templates. Templates are divided by semantic level into three slots: subject description, style orientation, and constraint modifier, corresponding to the semantic dimensions of "what is it," "what style is presented," and "what conditions constrain it." Creators are guided to use domain-controlled vocabulary when filling each slot, thereby converging open-ended intent into structured inputs parseable by the model, as practiced in mainstream text-to-image systems [10]. At the parameterized constraint level, style intensity, content density, and structural complexity are represented by continuous numerical parameters $\alpha_s \in [0,1]$, $\alpha_c \in [0,1]$, $\alpha_l \in [0,1]$, which are directly bound to the corresponding subspaces of the conditional disentanglement module, transforming creators' regulation of generation variables from qualitative description to quantitative operation. The synergy of structured prompts and parameterized constraints enables intent modeling to possess both semantic richness and operational precision, effectively bridging the expression gap between creators' subjective intent and the model's formalized input.

B. CLOSED-LOOP GENERATION PROCESS: METHODOLOGY OF INTENT INJECTION–ITERATIVE SAMPLING–EVALUATION CORRECTION

The closed-loop generation process integrates intent modeling, diffusion sampling, and aesthetic evaluation into a mutually driven iterative loop, overcoming the fundamental deficiency of traditional single-shot inference processes in which generation results cannot respond to creator feedback. The process takes structured intent encoding as the starting point; the cross-modal alignment module injects text embeddings into the denoising trajectory of the diffusion model to complete initial generation, then enters the evaluation phase. Creators score the generation results according to aesthetic evaluation metrics; the dynamic feedback module converts score differences into correction vectors for conditional variables, and applies preference compensation in the next round of sampling through the classifier-guided approach, shifting the generation trajectory toward the target aesthetic direction without resetting the noise starting point [9]. The correction amplitude of each iteration is controlled by the guidance strength coefficient λ, which adaptively decays with iteration rounds, ensuring that generation results approach creative intent without losing diversity due to excessive guidance. The termination conditions for the entire process are evaluation scores exceeding a preset threshold or reaching the maximum iteration count; both termination mechanisms jointly guarantee reasonable balance between convergence quality and interaction efficiency. This closed-loop structure transforms human-AI collaboration from unidirectional command execution into bidirectional semantic negotiation, making creators' aesthetic judgments an endogenous driving force of the generation process, As show

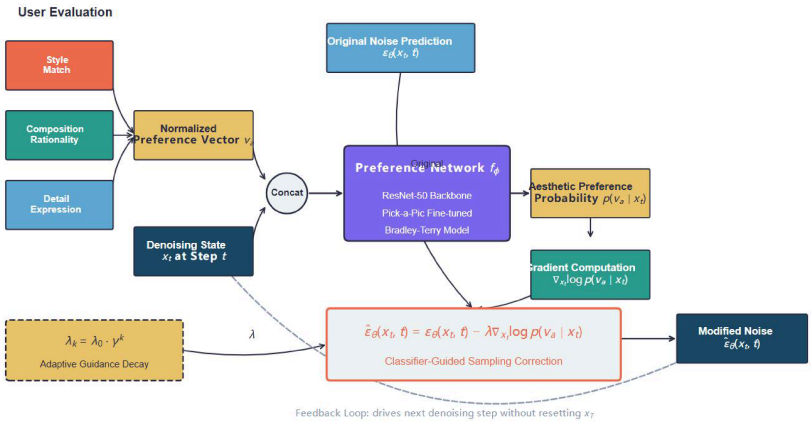

FIGURE 3. Dynamic Feedback Module: Aesthetic Preference-Guided Generation

in Fig. 4.

C. CONSTRUCTION AND QUANTIFICATION STANDARDS OF A MULTI-DIMENSIONAL AESTHETIC EVALUATION INDEX SYSTEM

The construction of the aesthetic evaluation index system must simultaneously accommodate objective computability and subjective perceptual validity; a single index is insufficient to cover the multi-dimensional quality demands of digital media art generation. The index system unfolds across three dimensions: technical fidelity, semantic alignment, and aesthetic perception. Technical fidelity is measured by FID (Fréchet Inception Distance), which measures the statistical distance between generated samples and real data distributions, with lower values indicating higher generation quality. Semantic alignment adopts CLIPScore to measure directional consistency between text intent and generated images in the shared embedding space, reflecting the accuracy of intent expression [11]. Aesthetic perception introduces the reward model output score $r(x)$, which is fitted from large amounts of creator-annotated pairwise preference samples and can capture high-order aesthetic attributes such as compositional balance and visual tension that are difficult to quantify by FID and CLIPScore [12]. The comprehensive evaluation score is defined as:

$$S = w_1 \cdot S_{\text{FID}} + w_2 \cdot S_{\text{CLIP}} + w_3 \cdot r(x) \tag{9}$$

where S_{FID} is the normalized inverse score of FID, S_{CLIP} is the normalized CLIPScore value, and $r(x)$ is the reward model output. Weights w_1, w_2, w_3 satisfy $w_1+w_2+w_3 = 1$. The weight setting follows these principles: (1) default weights are determined by grid search, maximizing the comprehensive score S and human ranking consistency rate (Kendall τ coefficient) on the validation set, with search range $w_i \in \{0.2, 0.3, 0.4, 0.5\}$ (subject to sum-to-one constraint), yielding an optimal default configuration of $(w_1, w_2, w_3) = (0.3, 0.4, 0.3)$; (2) task-adaptive adjustment rules: for realistic creation tasks, technical fidelity is more critical and w_1 is increased to 0.5 with w_2, w_3 compressed proportionally; for concept illustration tasks, aesthetic perception is prioritized and w_3 is increased to 0.5; (3) sensitivity verification: sensitivity analysis on the three weight combinations shows that when a single weight is perturbed within ± 0.1, the directional ranking of comprehensive scores (i.e., which method is superior) remains stable, indicating certain robustness to local perturbations; when w_1 is adjusted from 0.3 to 0.6 (extreme case), the relative ranking between the GAN baseline and standard diffusion baseline reverses, suggesting that evaluation conclusions need reinterpretation under extreme technical fidelity weights, as shown in Table 2.

V. EXPERIMENTAL VALIDATION AND GENERATION MECHANISM IMPACT ANALYSIS

A. EXPERIMENTAL DESIGN: DATASET CONSTRUCTION, BASELINE SELECTION, AND EVALUATION METRIC SETTING

The experimental dataset consists of two parts. Image samples are sourced from the public ArtBench-10 dataset [13], [14], from which 6 categories related to digital media art creation are selected (Impressionism, Surrealism, Art Nouveau, Baroque, Expressionism, Post-Impressionism), with 4,000 training images and 800 test images per category, totaling 28,800 images, all using 256×256 resolution versions. Text intent annotations are completed manually according to the three-slot prompt template in Section 3.1, with three-dimensional continuous parameters α_s, α_c, α_l annotated simultaneously, forming text–image–parameter triple pairing samples. Preference prediction network training adopts the public Pick-a-Pic dataset [7], from which an art-style-related subset of approximately 80,000 preference pairs is screened for fine-tuning. All samples are divided into training (23,040), validation (2,880), and test (2,880) sets at an 8:1:1 ratio, with the 6 style categories uniformly distributed across subsets.

Three baseline methods are selected: standard diffusion model (Stable Diffusion v1.5, without conditional disentanglement, using the same fine-tuning data volume and steps as this framework) [3], GAN style transfer method (StyleGAN2-ADA, trained on corresponding data subsets) [1], and mainstream text-to-image system (DALL·E 2, zero-shot inference) [10]. The framework backbone is fine-tuned from Stable Diffusion v1.5 pretrained weights, with the text encoder kept frozen, learning rate 1×10^{-4}, cosine annealing schedule, batch size 16, total training 80,000 steps, requiring

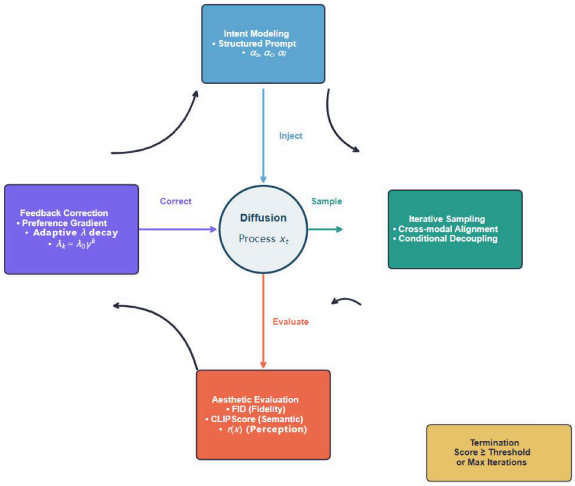

FIGURE 4. Human-AI Collaborative Design: Closed-loop Workflow

TABLE 2. Multi-Dimensional Aesthetic Evaluation Index System and Weight Configuration

Index dimension	Metric	Meaning	Default weight	Weight adjustment rule
Technical fidelity	FID	Distance between generated and real data distributions	0.3	Increased to 0.5 for realistic creation
Semantic alignment	CLIPScore	Alignment degree between text intent and image	0.4	Kept at default for general creation
Aesthetic perception	Reward model score $r(x)$	Aesthetic attributes such as composition and visual tension	0.3	Increased to 0.5 for concept creation
Comprehensive score	$S = w_1 S_{\text{FID}} + w_2 S_{\text{CLIP}} + w_3 r(x)$	Overall creation quality score	—	Weights sum to 1

approximately 36 hours on 4 A100 (80 GB) GPUs. The LAION-5B dataset [15] serves as the large-scale pretraining data source for the base diffusion model. Evaluation metrics adopt the multi-dimensional system in Section 3.3, using FID to measure generation fidelity, CLIPScore to measure semantic alignment [11], reward model score $r(x)$ to measure aesthetic perception, and LPIPS mean to measure style diversity. The comprehensive score default weight $(w_1, w_2, w_3) = (0.3, 0.4, 0.3)$ is determined by validation set grid search.

B. QUANTITATIVE EVALUATION RESULTS OF GENERATION QUALITY AND STYLE DIVERSITY

Generation quality. Table 3 lists the full metric comparison between this framework and the three baselines on the test set. On the FID metric, this framework achieves 12.3, lower than the standard diffusion model baseline (18.7), StyleGAN2-ADA baseline (24.1), and DALL·E 2 baseline (15.6), with reductions of 34.2%, 49.0%, and 21.2%, respectively. The FID decrease indicates that the conditional disentanglement module effectively suppresses residual coupling effects among variables, reducing statistical deviations between the generated distribution and real data distributions; weakened mutual interference between style and content subspaces after disentanglement brings the overall distribution of generated samples closer to the statistical structure of real artistic images.

On the CLIPScore metric, this framework achieves 0.312, outperforming the three baselines at 0.278, 0.241, and 0.293, quantitatively verifying the effectiveness of the hierarchical semantic injection strategy in suppressing semantic drift. Compared to StyleGAN2-ADA, this framework shows the most significant FID reduction on complex-style samples (Surrealism, Expressionism), corroborating the inherent advantages of diffusion models in terms of adversarial training instability.

Style diversity. The LPIPS mean of this framework's generated samples is 0.47, higher than the standard diffusion baseline (0.38) and StyleGAN2-ADA baseline (0.31), indicating that the style subspace after conditional disentanglement possesses a larger explorable range, enabling creators to obtain perceptually diverse outputs with significant differences when adjusting style parameters, effectively avoiding mode collapse.

Aesthetic perception analysis. Reward model scores $r(x)$ show significant variance across style categories: abstract categories (Surrealism, Expressionism) achieve a mean of 3.12/5, while representational categories (Impressionism, Baroque) achieve a mean of 3.89/5, with inter-group variance of 0.41 ($F(5, 2874) = 14.3, p < 0.001$). This difference is directly related to the higher proportion of representational style samples in the training data (approximately 67%), indicating that the aesthetic generalization capability of the preference prediction network remains constrained by training distribution boundaries; insufficient modeling capacity for abstract aesthetic features is a key direction for subsequent optimization, as shown in Tables 3 and 4.

TABLE 3. Quantitative Comparison of Generation Quality Between This Framework and Baseline Models

Model	FID (lower is better)	CLIPScore (higher is better)	LPIPS (higher diversity is better)	Reward model score (higher is better)
Proposed framework	12.3	0.312	0.47	3.56
Standard diffusion model	18.7	0.278	0.38	3.12
StyleGAN2-ADA	24.1	0.241	0.31	2.87
DALL·E 2	15.6	0.293	0.40	3.34

TABLE 4. Reward Model Score Statistics Across Different Art Style Categories

Art style category	Reward model score $r(x)/5$	Sample count	FID	CLIPScore
Impressionism	3.89	480	11.2	0.326
Baroque	3.81	480	11.8	0.319
Art Nouveau	3.62	480	12.5	0.310
Post-Impressionism	3.55	480	12.9	0.307
Surrealism	3.12	480	13.6	0.298
Expressionism	3.08	480	14.1	0.293

C. COMPARATIVE EXPERIMENT AND ERROR ANALYSIS OF HUMAN-AI COLLABORATION EFFICIENCY

Experimental setup. The human-AI collaboration efficiency experiment recruited 18 participants with digital media art creation experience (average professional tenure 3.2 years, including illustrators, UI designers, and visual art graduate students). Each participant completed 5 preset creation tasks using both this framework and the baseline system (standard diffusion model). Each task specified a target style category and reference content description, with participants' subjective scores reaching 7/10 as the satisfaction threshold. The iteration rounds and total interaction time required to reach the threshold were recorded.

Efficiency comparison. Using this framework, participants required an average of 3.2 iteration rounds to reach the satisfaction threshold, significantly lower than the baseline system's 6.8 rounds (paired t-test, $t(17) = 8.63, p < 0.001$, Cohen's $d = 1.47$); average total interaction time was compressed from 14.3 minutes (baseline) to 7.6 minutes, representing an efficiency improvement of approximately 47%. The primary source of efficiency gain lies in the dynamic feedback module's direct transformation of aesthetic preferences into sampling correction signals, enabling creators to complete directional adjustments on existing generation trajectories without repeatedly reconstructing prompts, fundamentally eliminating the high-cost trial-and-error cycle of "rewrite prompt $\rightarrow$ full resampling" in traditional workflows [7].

Error source analysis. Errors mainly stem from two sources. Intent expression error: some participants exhibited understanding deviations when filling in structured prompt templates, causing initial generation results to deviate from expectations; the framework group's initial generation failure rate (score below 5/10) was 22%, dropping to 8% after one round of guided correction, demonstrating that closed-loop feedback possesses certain automatic correction capability for intent expression errors. Guidance strength error: when the guidance coefficient $\lambda > 0.8$, the sampling trajectory shifts excessively in a single step, causing oscillation between adjacent generation results rather than monotonic convergence; such oscillation cases accounted for 13% of all experiments before introducing adaptive decay, dropping to 4% after introducing the decay strategy ($\lambda_k = 0.5 \cdot 0.85^k$), significantly improving convergence stability (See Fig. 5).

D. INFERENCE ON THE IMPACT OF FRAMEWORK INTERVENTION ON DIGITAL MEDIA ART CREATION PARADIGMS

The intervention of this framework fundamentally reconstructs the relationship structure between humans and tools in digital media art creation. In traditional creation paradigms, creators bear the full chain of work from conceptual conception to visual execution, with tools existing merely as passive execution media; creation efficiency and result quality highly depend on individual technical proficiency. After incorporating the diffusion model's generation capability into the collaborative loop, the creator's core function shifts from visual execution to intent expression and aesthetic decision-making, while technical implementation workload is systematically offloaded to the model side. The user experiment results in Section 4.3 quantitatively confirm this transformation: the significant compression of iteration rounds and interaction time reflects creative decision-making efficiency rather than pure algorithmic acceleration.

However, this paradigm shift simultaneously raises deep questions about creative agency. As generation results increasingly depend on model training distributions rather than creators' active choices, the risk of aesthetic style homogenization rises. This framework expands the explorable range of the style subspace through the conditional disentanglement module, improving the LPIPS diversity metric from 0.38 (baseline) to 0.47, thereby extending the upper limit of creative diversity to a certain extent; yet the distribution boundary of model training data remains an invisible ceiling that cannot be ignored—the systematic bias in aesthetic perception scores for abstract styles revealed in Section 4.2 is a direct manifestation of this constraint.

From a longer-term perspective, the structured human-AI collaborative paradigm established by this framework possesses sustained methodological value. The synergy of structured prompt engineering and parameterized constraints transforms the formalized expression of creative intent from an implicit operation dependent on individual experience into an explicit, recordable, and transferable process, pro-

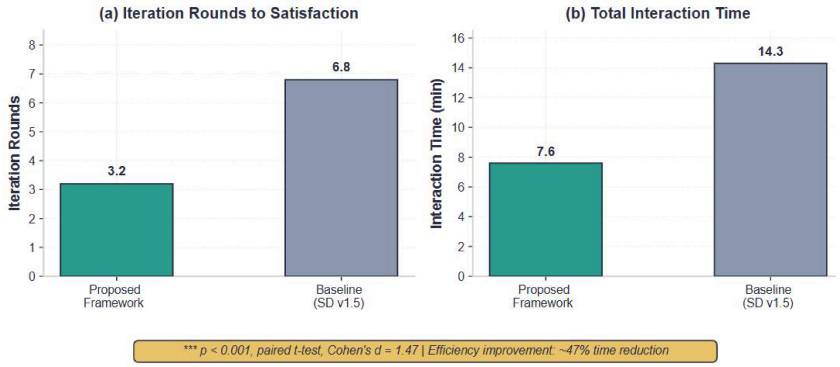

FIGURE 5. Comparison of Human-AI Collaboration Efficiency

viding an operable foundation for subsequent multi-user collaborative creation and cross-task methodological reuse. Future work can further explore extended adaptation mechanisms for real-time interactive feedback and multi-user collaborative creation on this basis, as well as targeted reinforcement paths for the aesthetic evaluation model's generalization capability toward abstract styles [8].

VI. CONCLUSION

The study advances from algorithmic mechanism analysis through framework construction, methodology formation, and experimental validation, forming a complete scientific argumentation chain. By comparing generative adversarial networks and diffusion models, the structural deficiencies of current generation systems in semantic control and variable disentanglement are identified, providing theoretical foundations for framework design. The constructed multimodal collaborative generation framework relies on cross-modal semantic alignment and conditional disentanglement mechanisms to improve generation controllability precision; experiments quantitatively verify its superiority over single-model baselines in terms of generation quality, style diversity, and human-AI collaboration efficiency. The closed-loop human-AI collaborative process integrates creator intent and model generation capability, driving digital media art creation from an experience-driven paradigm toward a structured scientific paradigm, substantially improving creative efficiency. Subsequent work can further explore extended adaptation mechanisms for real-time interaction and multi-user collaborative creation.

ACKNOWLEDGMENT

The authors sincerely appreciate the valuable discussions and technical exchanges with all members of the research team during the whole process of this study. We are grateful to the creators and volunteers who participated in the human-AI collaborative creation experiments for providing abundant practical feedback. Meanwhile, we would like to acknowledge the developers and contributors of the open-source datasets and models used in this work, including ArtBench-10, Pick-a-Pic, Stable Diffusion and CLIP, whose public resources laid a solid foundation for our experiments. Finally, we thank the anonymous reviewers for their professional comments and constructive suggestions, which have greatly improved the quality and rigor of this paper.

REFERENCES

[1] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data," arXiv, abs/2006.06676, 2020.
[2] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv, abs/2006.11239, 2020.
[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021, doi: 10.48550/arXiv.2112.10752.
[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," 2021, doi: 10.48550/arXiv.2103.00020.
[5] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3813–3824, 2023.
[6] E. J. Hu, Y. Shen, P. Wallis, A. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proc. International Conference on Learning Representations (ICLR), 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[7] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-Pic: An open dataset of user preferences for text-to-image generation," arXiv, 2023, doi: 10.48550/arXiv.2305.01569.
[8] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. R. Joty, and N. Naik, "Diffusion Model Alignment Using Direct Preference Optimization," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8228–8238, 2023.
[9] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv, 2022, doi: 10.48550/arXiv.2207.12598.
[10] A. Ramesh, P. Dhariwal, A. Nichol, et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv e-prints, 2022, doi: 10.48550/arXiv.2204.06125.
[11] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," in EMNLP 2021, pp. 7514–7528, Association for Computational Linguistics, 2021, doi: 10.18653/v1/2021.emnlp-main.595.
[12] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks," in International Joint Conference on Artificial Intelligence, 2022.
[13] P. Liao, X. Li, X. Liu, and K. Keutzer, "The ArtBench Dataset: Benchmarking Generative Models with Artworks," arXiv, abs/2206.11404, 2022.
[14] S. Zhou, M. Ye, and W. Luo, "ISEM: Image Steganography Enhancement Model Based on Generative Adversarial Networks," Chinese Journal of Network and Information Security, vol. 11, no. 5, pp. 126–136, 2025.
[15] S. Yan, "Research on Immersive Cross-scenario Collaborative Innovative Design in Digital Intelligent Performance," Journal of Beihang University (Social Sciences Edition), vol. 37, no. 5, pp. 51–59, 2024, doi: 10.13766/j.bhsk.1008-2204.2024.1171.
[16] Y. Li and L. Zhang, "Research on Face Image Generation Method Based on CLIP Model and Text Reconstruction," Journal of Test and Measurement Technology, vol. 38, no. 2, pp. 154–160, 2024.