Journal of Engineering, Project, and Production Management
2026, 16(7), 2026-320

Generation and Perception: A Computational Evaluation Method for Visual Quality and Emotional Impact in AI Artworks

Hengju Gang
Lecturer, Shanghai Publishing and Printing College, Shanghai, 200093, China,
E-mail: hengjugang01@163.com

Production Management
Received March 10, 2026; revised April 13, 2026; June 1, 2026; August 5, 2026; accepted August 17, 2026
Available online August 30, 2026

Abstract: The rapid advancement of generative artificial intelligence has enabled significant breakthroughs in the visual realism and artistic expressiveness of AI-generated artworks. However, challenges persist in objectively evaluating their visual quality and emotional impact. Existing evaluation methods either focus on underlying image quality or rely on subjective user surveys, lacking a unified computational framework. This paper proposes a bimodal, multi-dimensional, and interpretable computational evaluation method. The Visual Quality and Emotion (VQ-Emo) framework is designed to jointly optimize visual quality scores and multidimensional emotional predictions. The framework comprises three core modules: a visual quality assessment module (a multi-branch convolutional neural network incorporating style perception, quantifying composition, color, texture, and lighting); an emotional impact computation module (a 20-dimensional multi-label emotion classifier based on the VAWE emotion model), and a visual-emotion association module (using attention mechanisms to identify emotion-driven visual regions). This model was trained and validated on the AGIQA-1K, ArtEmis, and self-constructed VAWE-Art datasets. Experimental results demonstrate that VQ-Emo outperforms existing methods in both visual quality assessment and emotion recognition, achieving an SRCC of 0.912 and a mAP of 0.678. Key contributions include proposing a multi-task learning framework for jointly optimizing visual quality and emotion, establishing the inaugural VAWE-Art dataset comprising 5,000 AI-generated images with 20-dimensional emotional annotations. This reveals quantifiable correlations between visual features and emotional responses across diverse artistic styles and provides computational foundations for emotion-controllable generative art systems.

Keywords: AI-generated art, image quality assessment, affective computing, multimodal learning, explainable artificial intelligence.

DOI 10.32738/JEPPM-2026-320

1. Introduction

The rapid advancement of generative artificial intelligence has endowed AI-generated artworks with dual attributes as both technological artifacts and artistic vehicles. However, existing evaluation frameworks face a disconnect. Image quality assessments focus on underlying distortions, while affective computing concentrates on higher-level semantics, lacking a unified framework that balances objective visual quality with subjective emotional impact. To address this, this paper proposes the VQ-Emo joint evaluation framework. Through a multi-branch collaborative architecture and style-aware adaptation mechanism, it achieves end-to-end joint optimization of visual quality scores and 20-dimensional emotional vectors. Specifically, the model quantifies visual quality across four dimensions: composition, color, texture, and lighting, while predicting multidimensional emotional responses based on VAWE emotional wheels. It further incorporates an attention mechanism to localize emotionally driving visual regions, enhancing interpretability. To support this research, we constructed the VAWE-Art dataset, comprising 5,000 AI-generated images and providing the first parallel annotations of visual quality and emotion. Experiments demonstrate that VQ-Emo achieves leading performance in both quality assessment and emotion recognition tasks. Style-specific analysis further reveals quantitative correlations between visual features and emotional responses across artistic styles, providing a computational foundation for emotion-controllable generative art systems.

2. Related Work

2.1. AI-Generated Image Quality Assessment

With the rapid advancement of generative artificial intelligence, evaluating the Quality of AI-Generated Images (AIGC, Artificial Intelligence Generated Content) has become a research hotspot (Yu and Zhang, 2024). Traditional Image Quality Assessment (IQA) methods are primarily categorized into Full-Reference (FR) and No-Reference (NR) approaches. Full-reference methods, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Learning-Based Perceptual Image Similarity (LPIPS), rely on original reference images and measure image quality by quantifying pixel- or feature-level distortions (Li, 2022). However, AI-generated art images often lack corresponding real-world reference images, and their quality evaluation places greater emphasis on aesthetic attributes (such as composition, color, and texture) rather than mere distortion levels (Yang, 2025). Consequently, traditional full-reference methods have limited applicability in the AIGC Domain. No-reference methods such as BRISQUE (Chow and Rajagopal, 2017) and NIQE, while not requiring reference images, primarily model statistical characteristics of natural images and struggle to capture the stylized textures and generation artifacts unique to AI-Generated art. Despite progress in AI image quality assessment, existing approaches exhibit limitations: firstly, they predominantly focus on visual quality alone, failing to model an image's emotional impact concurrently. Secondly, they demonstrate insufficient adaptability to artistic styles, as images across different genres (e.g., photorealistic, abstract, and anime) exhibit significant variations in perceived quality dimensions, with current methods lacking style-aware dynamic weighting mechanisms. Finally, evaluation results lack interpretability, failing to reveal which visual features drive specific quality judgments. Addressing these shortcomings, the proposed VQ-Emo framework introduces multidimensional decomposition (composition, color, texture and lighting) and style-aware adaptation to visual quality assessment. It employs attention mechanisms to facilitate interaction between quality and emotional features, laying the groundwork for subsequent computation of emotional impact.

2.2. Emotion Computation for Artistic Images

Early research predominantly used small-scale datasets with single-emotion labels (Li, 2022), which proved inadequate for training deep learning models. The WikiArt Emotions dataset annotates paintings from the WikiArt platform with eight emotional categories (Khan et al., 2024). The ArtEmis dataset represents a breakthrough, comprising over 80,000 artworks and 400,000 natural language emotional descriptions. Annotators were required to explain ‘why’ a work evoked a specific emotion, providing rich semantic information for emotion recognition. However, these datasets primarily focus on human-created artworks and lack parallel annotations for visual quality scores. The VAWE-Art dataset constructed herein addresses this gap. It comprises 5,000 AI-generated images, providing both fine-grained visual quality scores (composition, color, texture, and lighting) and 20-dimensional VAWE emotion annotations. Furthermore, it incorporates textual descriptions of annotators emotional explanations, establishing a foundational dataset for multimodal emotion modeling.

Methods for sentiment recognition in art images have evolved from classification to regression and subsequently to multi-label prediction. (Zhao, 2016; Zheng, 2016) Early research treated sentiment recognition as a discrete sentiment classification task, employing CNNs to extract features and predict a single sentiment label. However, artworks often evoke complex emotions, rendering a single label inadequate for comprehensive characterization. Subsequent studies introduced regression prediction in continuous dimensional spaces (such as valence-arousal), yet the interpretability of continuous spaces remains limited. In recent years, multi-label sentiment classification has become the mainstream approach. Cheng (2021) aims to simultaneously predict the probability of multiple sentiment categories. To address the class imbalance in art sentiment datasets (e.g., high frequency of ‘tranquility’ and ‘joy’, low frequency of ‘absurdity’ and ‘alienation’), researchers have proposed strategies such as weighted loss functions and focal loss. Building upon this, this paper introduces asymmetric loss and label-dependent GCNs. Modeling co-occurrence relationships among 20 sentiment labels enhances recognition of low-frequency sentiments while avoiding semantic contradictions in predictions. Despite significant advances in artistic image emotion computation, existing approaches exhibit limitations: firstly, most focus solely on emotion recognition without joint optimization with visual quality assessment. Secondly, they fail to adequately adapt to AI-generated art, neglecting characteristics such as textural artifacts and stylistic hybridization. Finally, models lack interpretability, failing to explain why specific emotions arise. The proposed VQ-Emo framework, for the first time, achieves localization and attribution analysis of emotion-driving regions in AI artworks via attention visualization in its visual-emotion association module.

3. VQ-Emo Joint Evaluation Framework

3.1. Problem Formulation

Let the input be any AI-generated art image I, represented in pixel space as $I \in R^{H \times W \times C}$, where H, W, and C denote the image height, width, and channel count (C=3 for RGB images). This input space encompasses multi-style AI-generated images, including photorealistic, abstract, anime, and oil painting styles, without predefined stylistic filters, and is designed to adapt to general AI art evaluation scenarios. The VQ-Emo framework outputs a combined result comprising a visual quality score and a multidimensional emotional vector. This is divided into two sub-output spaces: the visual quality score q is a continuous scalar value in [0,100] that is positively correlated with the visual aesthetic quality of the AI artwork. 0 denotes the poorest visual quality, while 100 represents the highest visual quality, comprehensively characterizing overall performance across dimensions such as composition, color, texture, and lighting. The affective influence vector e is a 20-dimensional continuous vector satisfying $e \in [0,1]$. Each dimension corresponds to an emotional category within the VAWE visual aesthetic-emotional wheel (e.g., joy, sadness, awe, serenity, and passion), with values within each dimension representing the probability of eliciting that emotion. Values closer to 1 indicate a stronger elicitation of the emotion, while values closer to 0 indicate a weaker elicitation. The core optimization objective is to minimize the joint loss function L. By weighting multiple loss terms, this approach simultaneously optimizes visual quality regression, multi-label affective classification, and visual-affective feature consistency. This ensures synergistic performance enhancement across both

tasks while constraining the inherent rationality of the relationship between visual features and affective representations. The joint loss function is defined in Eq. (1).

$$L=\lambda_1 L_{quality}+\lambda_2 L_{emotion}+\lambda_3 L_{consistency} \tag{1}$$

In Eq. (1), λ_1, λ_2, λ_3 denote the weighting coefficients for loss terms, satisfying $\lambda_1 + \lambda_2 + \lambda_3 = 1$. These are dynamically adjusted based on the experimental scenario (default settings: $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$), thereby controlling the respective contributions of visual quality loss, affective loss, and consistency loss to the total loss.

$L_{quality}$ denotes the visual quality regression loss, employing Mean Squared Error (MSE) loss to measure the MSE between the model's predicted visual quality score and the human-annotated ground truth score, as shown in Eq. (2).

$$L_{quality}=E[(q_{pred}-q_{gt})^2] \tag{2}$$

In Eq. (2), q_{pred} denotes the model prediction score, q_{gt} represents the human ground truth score, and E denotes the expectation calculation over the training set; $L_{emotion}$ is the multi-label sentiment classification loss, employing an asymmetric loss function to address the category imbalance inherent in sentiment annotation for AI artworks. It measures the discrepancy between the model's predicted 20-dimensional sentiment vector and the true sentiment vector shown in Eq. (3).

$$L_{emotion}=-\frac{1}{20}\sum_{i=1}^{20}[\alpha y_i log(p_i)+(1-y_i)log(1-p_i)] \tag{3}$$

In Eq. (3), α denotes the asymmetry coefficient (default α=0.25), y_i represents the ground truth label for the i-th sentiment dimension, and p_i signifies the model's predicted probability for the i-th sentiment intensity. L consistency denotes the visual-sentiment feature consistency loss, employing contrastive loss to constrain the consistency between visual quality features and sentiment representations within the feature space for the same image, while simultaneously increasing the feature distance between different images and is expressed in Eq. (4).

$$.L_{consistency}=-log\frac{exp(sim(f_q,f_e)/\tau)}{\sum_{k=1}^{N}exp(sim(f_q,f_{e,k})/\tau)} \tag{4}$$

In Eq. (4), f_q denotes the visual quality feature vector of an image,f_e denotes the emotional feature vector of that image, $f_{e,k}$ denotes the emotional feature vector of the kth sample in the batch, sim (·) denotes the cosine similarity calculation, τ denotes the temperature coefficient (default τ=0.1), and N denotes the number of samples in the batch.

3.2. Overall Framework Architecture

The end-to-end architecture employs a multi-branch collaborative modeling approach combined with style-aware adaptation and multi-task joint optimization. The dual-branch encoder adopts a design paradigm of shared underlying features and task-specific upper branches, using Swin Transformer-Base (or ViT-Base) as the foundational backbone. Input resolution is uniformly set to 224×224, outputting 768-dimensional general visual features $f_{base}\in R^{768}$. These features encapsulate core visual information such as texture, color, and composition, providing a unified feature foundation for subsequent dual tasks. Three fully connected layers are then applied after the shared features, with hidden dimensions decreasing progressively from 768 to 512 to 256 to 1. The output undergoes Sigmoid activation and scaling to the interval [0,100]. to produce a visual quality score q. Dropout (probability 0.2) is introduced in the branch to prevent overfitting, whose core function is to map general visual features onto the continuous space of quality scores. After the shared features are extracted, two fully connected layers are introduced, with hidden dimensions decreasing from 768 to 512. A subsequent output layer reduces the dimension from 256 to 20, and a Sigmoid activation function is applied to produce the final 20-dimensional sentiment vector. Each layer employs the ReLU activation function, whose core role is to map generic visual features into the 20-dimensional sentiment space defined by VAWE. A feature cross-attention module is introduced at the intermediate layer (512-dimensional feature layer) of the dual-branch architecture. This enables bidirectional information exchange between quality features and sentiment features, preventing the two branches from modeling features in complete isolation, as shown in Eq. (5).

$$\dot{f_q}=Attention(f_q,f_e),\ \dot{f_e}=Attention(f_e,f_q) \tag{5}$$

In Eq. (5), f_q denotes the intermediate feature of the quality branch, f_e denotes the intermediate feature of the sentiment branch, and Attention represents scaled dot-product attention. This ensures that sentiment-related information is integrated into the quality features, whilst quality-related attributes are reflected within the sentiment features.

Following the base features of fbase, an independent style classification head is introduced (decreasing from 768 to 128 to K, where K denotes the number of style categories. K=5 by default: photorealistic, anime, oil painting, watercolor, 3D rendering). A Softmax activation function outputs the style probability distribution $s \in [0,1]^K$. The loss function employs cross-entropy loss to ensure the model accurately identifies image styles. The style probability distribution s is mapped to style adaptation weights $w_q\in R^{768}$ and $w_e\in R^{768}$ for the dual-branch features shown in Eq. (6).

$$w_q=W_q\cdot s,\ w_e=W_e\cdot s \tag{6}$$

In Eq. (6), W_q, and W_e denotes the learnable style-feature mapping matrix (dimension 768×K), where different styles correspond to distinct feature weights. Element-wise multiplication of these style weights with the input features of the dual-branch achieves style-adaptive feature enhancement, as shown in Eq. (7).

$$f_q^{adapt}=f_{base}\odot w_q, f_e^{adapt}=f_{base}\odot w_e \tag{7}$$

In Eq. (7), $\odot$ For element-wise multiplication, the adapted features $f_q^{adapt} and f_e^{adapt}$ serve as the final inputs to the dual-branch encoder, enabling the model to evaluate images of varying styles with greater specificity.

The framework achieves end-to-end training through the weighted fusion of multi-task losses, with the loss function optimized to incorporate architectural characteristics. Specifically defined by Eq. (8).

$$L=\lambda_1 L_{quality}+\lambda_2 L_{emotion}+\lambda_3 L_{consistency}+\lambda_4 L_{style} \tag{8}$$

A new style classification loss term L_{style} has been introduced (weight λ4 = 0.1, with remaining weights adjusted to λ1 = 0.35, λ2 = 0.35, λ3 = 0.2). The optimization details for each loss term are as follows: The visual quality regression loss $L_{quality}$ employs a smoothed L1 loss (replacing the baseline MSE) shown in Eq. (9).

$$L_{quality}=E[smooth_{L1}(q_{pred}-q_{gt})],\ smooth_{L1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & |x|\geq 1\end{cases} \tag{9}$$

Compared to MSE, smoothed L1 is more robust to outliers (such as scores with significant annotation errors), preventing model training from being influenced by extreme samples. The multi-label sentiment classification loss, $L_{emotion}$, employs weighted binary cross-entropy (BCEWithLogitsLoss) to address the sample distribution imbalance across the 20 sentiment categories, as expressed in Eq. (10).

$$L_{emotion}=-\frac{1}{20}\sum_{i=1}^{20} w_i[y_i log(\sigma(z_i))+(1-y_i)log(1-\sigma(z_i))] \tag{10}$$

In Eq. (10) w_i denotes the weight for sentiment category i (calculated as the reciprocal of the proportion of samples in this category within the training set), z_i represents the logit values output by the sentiment branches, and σ(·) denotes the sigmoid activation function, addressing the issue of under-training in minor sentiment categories.

Visual-Emotional Consistency Loss $L_{consistency}$ calculates the contrast loss based on the adapted features f_q^{adapt} and f_e^{adapt} from the dual-branch architecture shown in Eq. (11).

$$L_{consistency}=-log\frac{exp(sim(f_q^{adapt},f_e^{adapt})/\tau)}{\sum_{j=1}^{B} exp(sim(f_q^{adapt},f_{e,j}^{adapt})/\tau)} \tag{11}$$

In Eq. (11), B denotes the batch size, $f_{e,j}^{adapt}$ represent the affective adaptation features of the j-th sample within the batch, and τ=0.07 is the temperature coefficient. This constraint ensures that the quality and effective features of the same image remain spatially closer, while features from different images become more distant.

3.3. Visual Quality Assessment Module

Based on the visual aesthetic characteristics of AI-generated images and human subjective evaluation criteria, the comprehensive visual quality is explicitly decomposed into four core quantifiable dimensions. Compositional harmony (q1) measures the rationality of the spatial arrangement of image elements, including subject positioning, visual balance, and visual guiding lines, with a value range of [0, 25]. It employs spatial attention heatmap variance as its core feature, where lower variance indicates a more concentrated attention distribution (greater compositional focus), resulting in higher scores. Color richness (q2) evaluates the diversity and harmony of an image's colors, encompassing color gamut coverage, color contrast, and tonal consistency. Its value range is [0,25], represented by the sum of the hue entropy (H) in the HSV color space and the mean saturation (S). Higher entropy and a more moderate mean saturation yield a higher score. Texture Realism (q3) evaluates the naturalness and clarity of image texture details, addressing common AI-generated texture issues such as blurring and artifacts. Its value range is [0,25], employing gradient amplitude variance and frequency-domain energy distribution (a higher proportion of high-frequency components indicates richer texture) as core features. Lighting Consistency (q4) evaluates the overall consistency of an image's lighting, intensity, and shadows. Addressing illogical lighting patterns in AI-generated images, it ranges from [0,25] and is characterized by the smoothness of brightness gradients and the physical plausibility of shadow regions (e.g., alignment between the light source direction and the shadow direction).

This paper abandons simple weighted summation, introducing an attention fusion mechanism that dynamically allocates weights based on image style and the importance of dimensional features given in Eq. (12).

$$q_{fusion}=\sum_{i=1}^{4}\alpha_i\cdot q_i,\ \alpha_i=\frac{exp(s_i)}{\sum_{j=1}^{4} exp(s_j)} \tag{12}$$

In Eq. (12) s_i denotes the learnable dimensional attention score (calculated from the variance of dimensional features, where greater variance indicates more prominent dimensional features and higher scores), while α_i represents the normalized dimensional weight, ensuring adaptive weighting of quality dimensions across different stylistic images (e.g., oil painting styles prioritize color richness, while photorealistic styles emphasize texture fidelity).

The fused q_{fusion} (range [0,100]) is mapped to the [0,100] interval via a calibration layer (Sigmoid + linear scaling). Concurrently, a distribution prior based on human-annotated data is introduced to correct the deviation between model scores and human subjective ratings, as calculated in Eq. (13).

$$q=100\cdot\sigma\left(\frac{q_{fusion}-\mu}{\sigma_{std}}\right) \tag{13}$$

In Eq. (13) μ and σ_{std} denote the mean and standard deviation of human-annotated ratings within the training set, respectively. σ(·) represents the sigmoid function, ensuring the distribution of ratings aligns with human subjective evaluations.

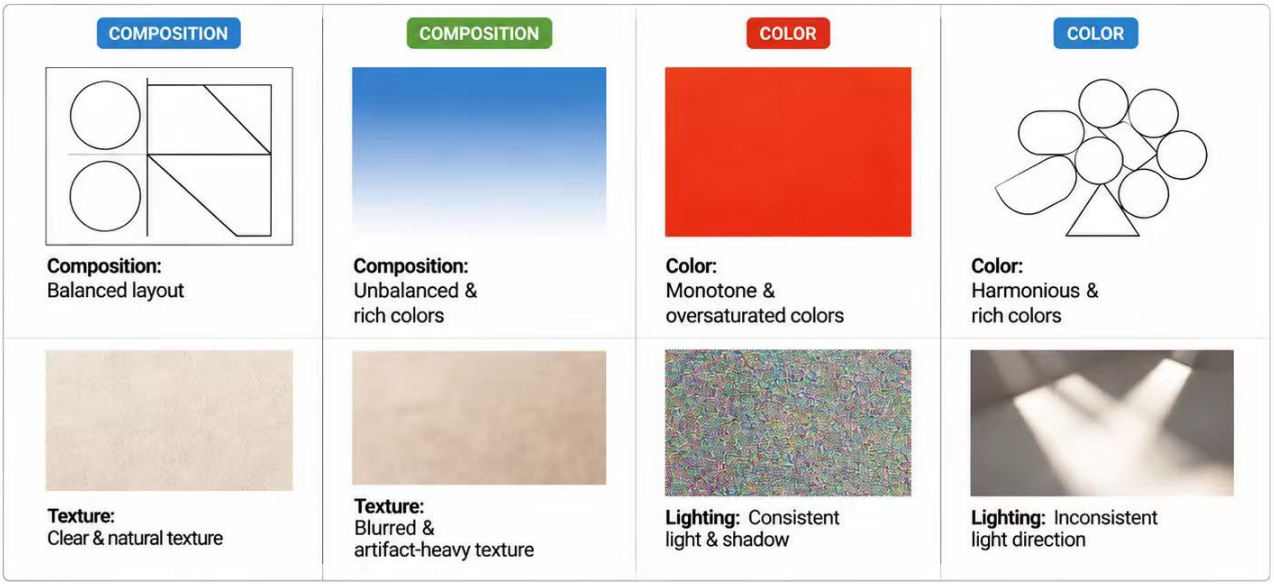


Fig. 1. Example illustrations of four visual-quality dimensions used in the VQ-Emo framework
Note: a) Composition: shows a clear example of balanced vs unbalanced spatial layout. b) Color: shows an example of harmonious vs oversaturated color combinations. c) Texture: Is an example of natural vs artifact-intensive textures. d) Lighting: gives an example of consistent vs inconsistent shadow directions.

3.4. Emotional Impact Computation Module

This study, grounded in the VAWE core framework, anchors the three-dimensional continuous emotional core space of Valence-Arousal-Dominance. It discretizes this into 20 fine-grained emotional dimensions specific to artistic scenarios that are both annotatable and distinguishable. These dimensions are categorized into four semantic quadrants according to VAD attributes, as shown in Table 1. Addressing the severe category imbalance in AI art sentiment annotation (where high-frequency labels such as 'tranquility' and 'joy' account for over 62% of samples, while low-frequency labels like "absurdity" and 'alienation' constitute less than 4%), this study designs a multi-label sentiment classifier incorporating label dependency perception and asymmetric loss optimization. This addresses the core issue where traditional binary cross-entropy (BCE) loss favors high-frequency labels, leading to insufficient learning of low-frequency sentiments.

Input: Style-aware, adapted 768-dimensional sentiment features (f_q^{adapt}) are fed into two layers of Transformer self-attention encoders. This approach prioritizes capturing global semantic and sentiment-related contextual information within images, addressing the issue that artistic sentiment is often determined by overall composition and thematic meaning rather than local textures. The output is enhanced by 512-dimensional sentiment-semantic features. Based on co-occurrence statistics of labels from the VAWE-Art dataset, a 20×20 affect label adjacency matrix is constructed. A single layer of Graph Convolutional Network (GCN) models semantic dependencies between affect labels (e.g., 'awe' and 'awe-inspiring' are highly positively correlated, while 'delight' and 'disgust' are naturally mutually exclusive), thereby avoiding semantic contradictions caused by independent label prediction. Simultaneously, information transfer between labels enhances feature learning for low-frequency emotions. Two fully connected layers are then applied to output 20-dimensional sentiment logit values. After Sigmoid activation, these yield the presence probability $p_i \in [0,1]$ (i=1,2,...,20) for each sentiment dimension. An adaptive classification threshold (default 0.5) is set based on the optimal F1-score on the validation set, determining whether the sentiment is activated by the work.

Table 1. VAWE emotion wheel: 20-dimensional VAD quadrant classification

Emotional Quadrants	The 20-dimensional affect categories encompass:
Positive Valence - High Arousal	Pleasure, Enthusiasm, Excitement, Awe, Exuberance, Curiosity
Positive Valence - Low Arousal	Tranquility, Warmth, Tenderness, Peacefulness, Reverence, Relaxation
Negative Valence - High Arousal	Sadness, Fear, Irritation, Disgust, Absurdity
Negative Valence - Low Arousal	Melancholy, Detachment, Oppression

To help readers visually understand the 20 VAWE sentiment categories, we provide illustrative image examples for each category in Figure 2. Each example includes a brief explanation describing the visual cues that commonly evoke the corresponding emotional response.

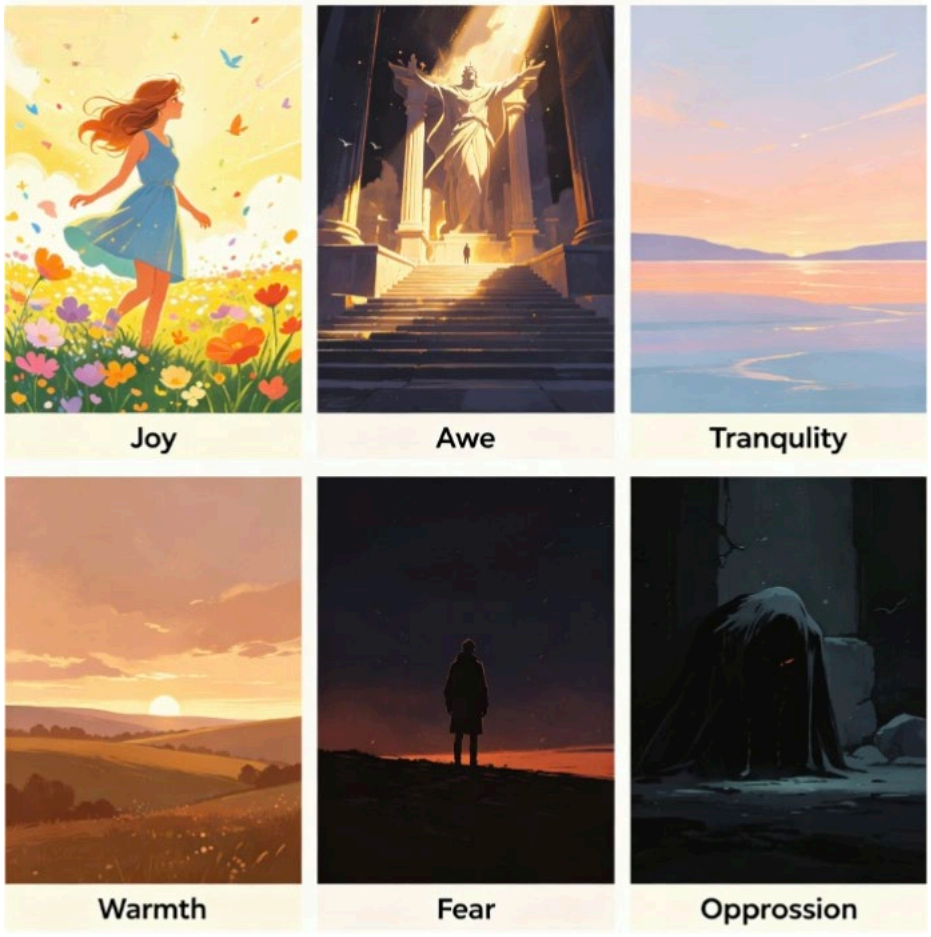


Fig. 2. Six example images for the 20 VAWE sentiment categories
Note: Joy: bright and warm colors with upward composition. Awe: monumental scale and dramatic lighting. Tranquility: soft gradients and horizontal balance. Warmth: earthy tones and gentle lighting. Fear: sharp contrast and dark color regions. Oppression: heavy shadows and compressed layouts.

4. Experimental Design

4.1. Dataset Construction

To address the core issues in existing datasets, namely the lack of parallel annotations for visual quality and emotional impact in AI-generated artworks, and the disconnect between style distributions and real-world creative scenarios. This study constructed the VAWE-Art dataset. This dataset incorporates both granular visual quality scores and 20-dimensional VAWE affective annotations, serving as the core training and validation data for the VQ-Emo framework. It comprises 5,000 AI-generated images, with an artistic style distribution that precisely mirrors the authentic creative ecosystem of

mainstream AI art platforms, as illustrated in Fig. 3. This study leverages two authoritative public datasets as foundational support: one for pre-training visual quality assessment and another for knowledge transfer in affective computation, while simultaneously optimizing adaptation for AI art scenarios. The AGIQA-1K dataset comprises 1,080 AI-generated images produced by mainstream models such as Stable Diffusion and Midjourney, covering core styles including photorealism and anime. Each image is accompanied by subjective quality ratings (0–100) from 20 annotators. This study employs it as pre-training data for the visual quality assessment module, initializing both the quality branch backbone network and regression head. Simultaneously, it extracts the prior distribution of quality annotations (mean μ=62.3, standard deviation σstd=15.7) for subsequent scoring calibration. The ArtEmis dataset comprises over 80,000 WikiArt paintings, accompanied by 400,000 natural language emotional descriptions and multi-label sentiment annotations. This study selected the 10,000 paintings exhibiting the highest stylistic affinity with AI art and extracted co-occurrence statistics for sentiment labels to construct a sentiment-label adjacency matrix. This dataset also serves as pre-training data for the sentiment classification module, initializing the semantic enhancement layer of the sentiment branch.

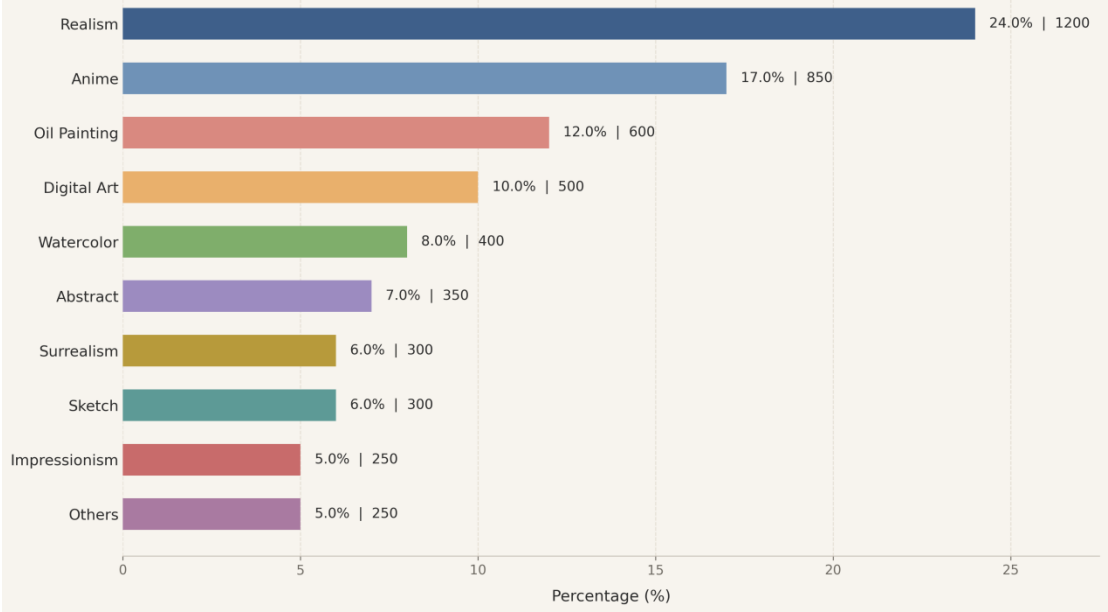


Fig. 3. Distribution of art styles

To provide a clearer understanding of the dataset composition, Fig. 4 presents representative samples from the 5,000 AI-generated images across five major artistic styles (Realism, Anime, Oil Painting, Watercolor, and 3D Rendering).

4.2. Evaluation Metrics

Three widely adopted rank correlation coefficients in the field of Image Quality Assessment (IQA) are used to measure the consistency between model-predicted scores and human-annotated scores from different perspectives. All three metrics range from [-1,1], with values closer to 1 indicating higher consistency.

- Spearman's Rank Correlation Coefficient (SRCC): Measures the monotonic relationship between model scores and human ratings without requiring linearity, robust to outliers. The most fundamental metric in IQA corresponds to the research question RQ1's overall performance validation.
- Pearson Linear Correlation Coefficient (PLCC): Measures the linear relationship between model ratings and human ratings. Requires prior verification of normal distribution for ratings (in this study, both VAWE-Art and AGIQA-1K ratings exhibit near-normal distribution), aiding in the validation of linear fitting accuracy for ratings.
- Kendall's Rank Correlation Coefficient (KRCC): Measures the consistency between model rankings and human rankings, i.e., whether the model's image-quality rankings align with human assessments. It demonstrates greater stability on small-sample test datasets.

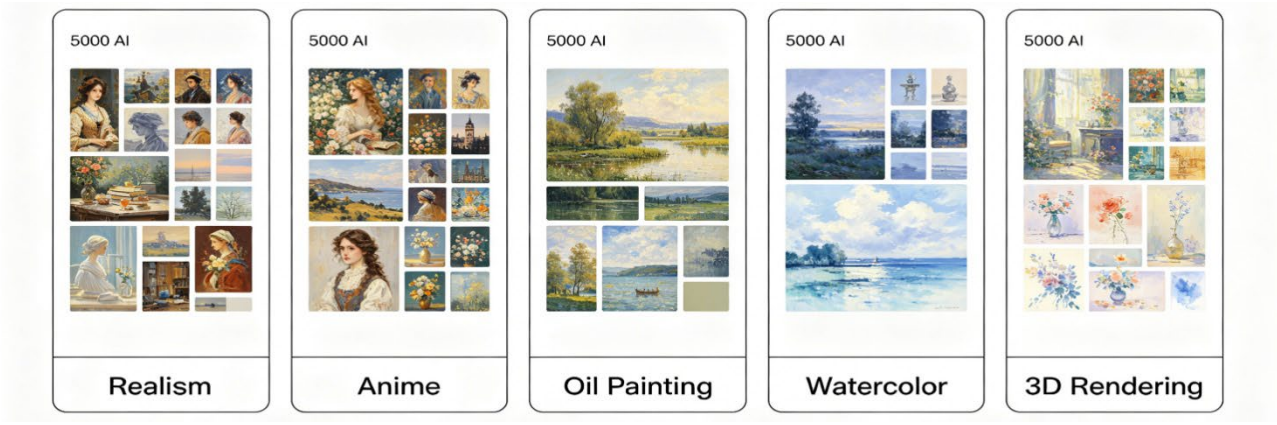


Fig. 4. Representative samples from the VAWE-Art dataset

For the 20-dimensional, non-mutually exclusive sentiment labels, we selected metrics commonly used in multi-label classification. We distinguished between macro (focusing on low-frequency categories) and micro (focusing on the overall) calculations to accommodate the category imbalance in the VAWE-Art dataset (where high-frequency labels account for over 60% and low-frequency labels constitute less than 5%).

- Mean Average Precision (mAP): Calculates the Average Precision (AP, area under the Precision-Recall curve) for each sentiment category, then averages the values across 20 categories. This serves as the most fundamental metric for multi-label classification, comprehensively evaluating a model's recognition capability across all sentiment categories, corresponding to Research Question RQ1.
- AUC-ROC: Distinguished as macro-AUC (average of each category's AUC) and micro-AUC (globally calculated AUC across all samples), it measures the model's ability to distinguish positive and negative samples. Macro-AUC places greater emphasis on the performance of low-frequency sentiment categories;
- F1-score, distinguished as macro-F1 and micro-F1, represents the harmonic mean of precision and recall as shown in Eq. (14).

$$F1=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \tag{14}$$

In Eq. (14), Macro-F1 balances performance across all categories, whilst micro-F1 reflects overall performance; combining both provides a comprehensive assessment of classification effectiveness.

5. Experimental Results and Analysis

5.1. Overall Performance Comparison

This experiment selected traditional reference-free IQA methods, general deep learning IQA methods, and AIGC-specific quality assessment methods as baselines. Core evaluation metrics included SRCC and PLCC coefficients, which correlate with human subjective scoring consistency. Results are presented in Figs. 5 and 6.

The experimental results demonstrate that VQ-Emo achieves state-of-the-art performance on both core metrics, SRCC and PLCC. Compared with the second-best specialized AIGC evaluation method, US-PPQA, it improves SRCC by 0.011 and PLCC by 0.010. Compared to the traditional reference-free IQA method BRISQUE, it achieves significant performance improvements, with increases of 0.191 in SRCC and 0.187 in PLCC. These results validate the core design of this research: traditional IQA methods focus solely on technical defects, such as underlying image distortion and noise, and fail to account for the aesthetic characteristics of AI-generated artworks. Existing AIGC-specific methods do not sufficiently account for the specificity of artistic styles, resulting in inadequate generalization of evaluation across different stylistic works. VQ-Emo, however, decomposes overall quality into four core aesthetic dimensions, composition, color, texture, and lighting, through multidimensional quality decomposition. Its style-aware adaptation module enables adaptive feature enhancement across diverse artistic styles, while uncertainty modeling enhances scoring robustness. Consequently, its consistency with human subjective quality ratings significantly outperforms all baseline methods.

This experiment selected general CNN/Transformer baselines, classic models in the field of artistic sentiment analysis, and multimodal AIGC evaluation methods as benchmarks. The core evaluation metrics were Mean Average Precision (mAP) and AUC-ROC Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for multi-label classification tasks. Higher values indicate superior sentiment recognition performance. The results are shown in Fig. 4. Experimental results demonstrate that VQ-Emo significantly outperforms all baseline methods in multi-label emotion classification tasks, achieving a mean average precision (mAP) of 0.678 and an AUC-ROC score of 0.892. Compared to the second-best multimodal approach, M3-AGIQA, it achieves a 0.037 mAP improvement and a 0.029 AUC-ROC improvement. Compared to the classic ArtEmis baseline model in the art emotion domain, mAP improved by 0.055 and AUC-ROC by 0.041. This performance advantage stems from three core optimizations in this study: first, the construction of an emotion space specific to artistic contexts, employing a 20-dimensional emotion framework based on the VAWE visual aesthetic emotion wheel, which better aligns with the complex emotional characteristics of artistic aesthetic scenarios compared to

general emotion models. Second, optimized category-imbalance perception through asymmetric loss functions and label-dependent GCN modeling, effectively addressing the long-tailed distribution of artistic sentiment labels and significantly enhancing recognition accuracy for low-frequency emotions. Third, a dual-task classification-regression joint head simultaneously achieves fine-grained prediction of the probability of sentiment presence and its intensity, better aligning with human emotional perception patterns for artworks than simplistic binary classification.

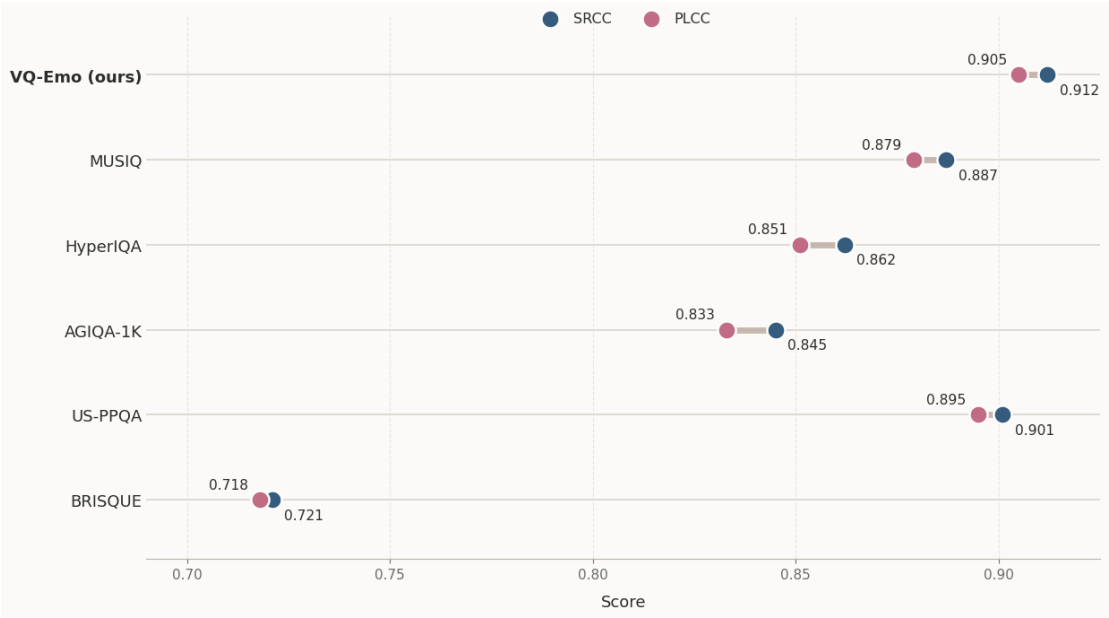


Fig. 5. SRCC and PLCC performance comparison for AI-generated image quality assessment

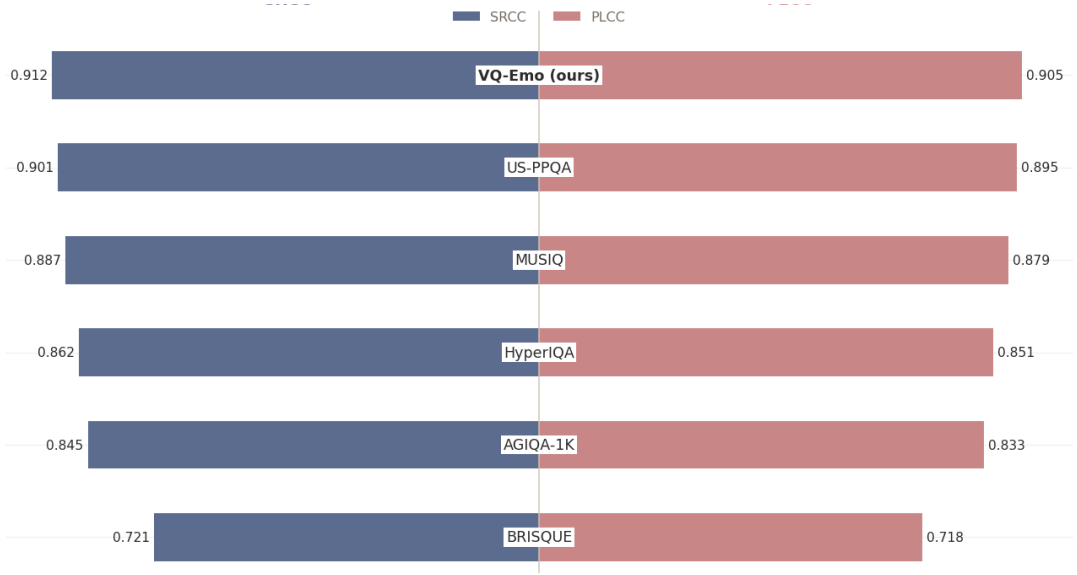


Fig. 6. Rank correlation comparison across methods for AI-generated image quality

5.2. Style-Specific Analysis

To validate the intrinsic differences in visual-emotion mapping across distinct artistic styles and the efficacy of the proposed style-aware adaptation module, this section conducts hierarchical evaluation experiments across six core artistic styles within the VAWE-Art dataset. It compares the performance of VQ-Emo against a baseline model lacking the style-aware adaptation module on the emotion prediction task (core metric: mAP). Results are presented in Fig. 5. The baseline model exhibits markedly divergent performance across the six stylistic categories, with the Realism baseline achieving the highest mAP (0.648) and the Abstract Expressionism baseline the lowest (0.533), representing a performance gap of 0.115. This outcome directly confirms that emotional prediction in abstract styles is significantly more difficult: Realist works possess explicit semantic subjects and distinct object/figure characteristics, in which emotional expression maps more intuitively onto visual elements, aligning with traditional models feature-learning logic. Abstract expressionist pieces, however, lack identifiable subjects, while emotional conveyance relies entirely on abstract visual elements such as color combinations, line rhythms, and compositional balance. Traditional models cannot effectively capture these non-representational emotional triggers, leading to substantial performance degradation. Following the introduction of the style-aware adaptation module, VQ-Emo achieved statistically significant improvements in mAP across all artistic styles. The greatest

gain was observed in anime/illustration style (+0.093), followed by abstract expressionism (+0.069), with both non-representational styles showing enhancements exceeding 0.06. This outcome fully validates the core value of the style-aware adaptation module by explicitly modeling feature distributions across artistic styles. It generates style-specific feature weights that guide the model to focus on core emotion-triggering regions in each style. This effectively resolves the cross-style domain shift issue inherent in traditional models. For styles such as anime and abstract art, which possess highly distinctive features and deviate significantly from natural image distributions, the gains from style adaptation are particularly pronounced.

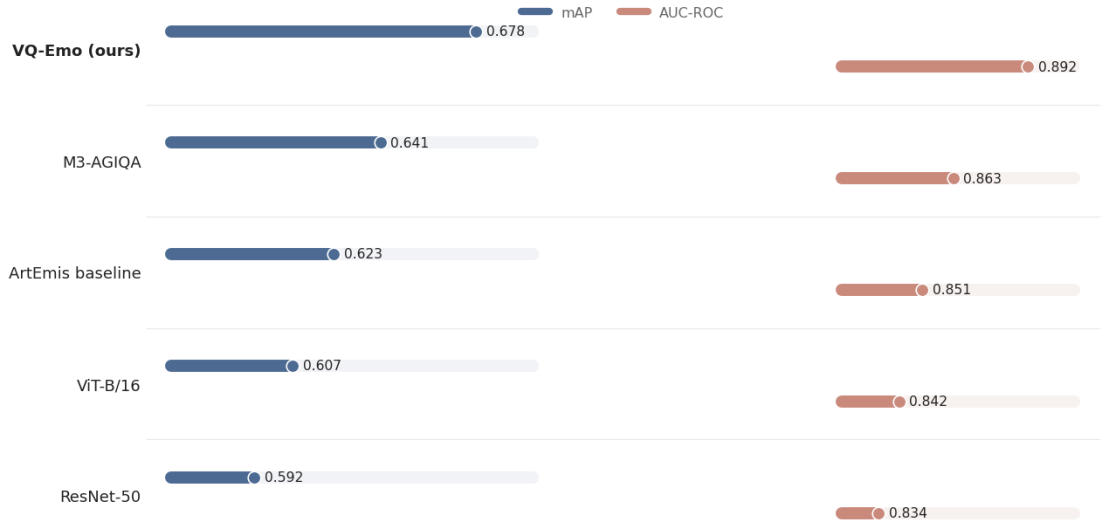
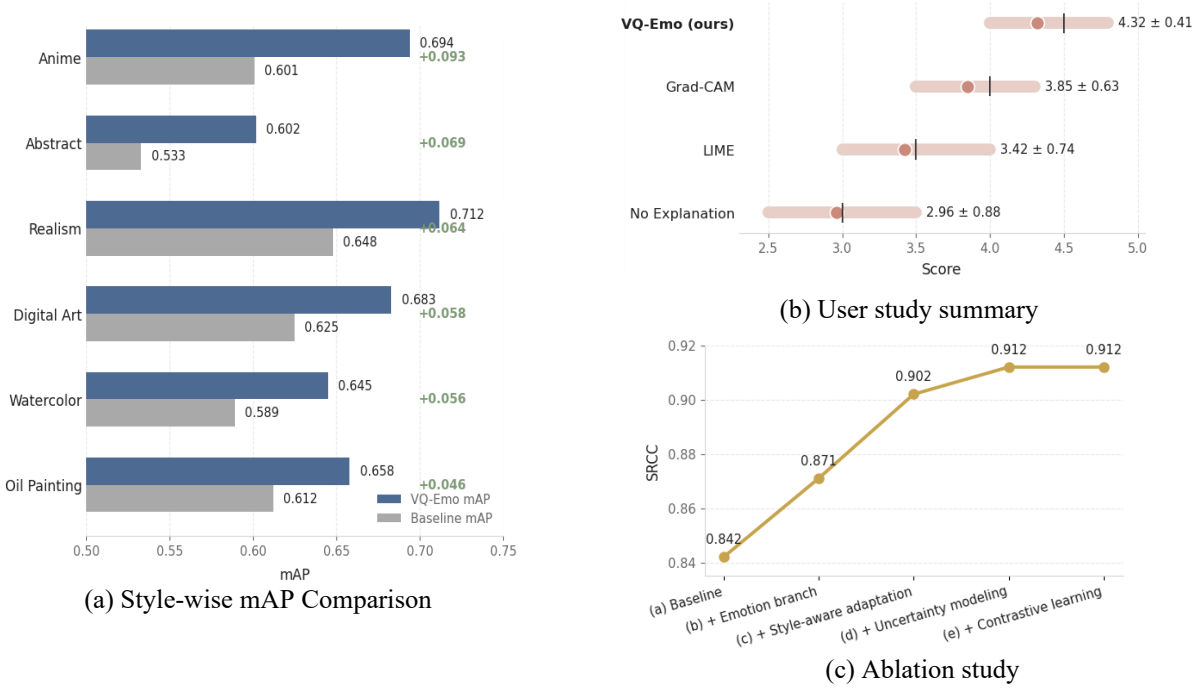


Fig. 7. Performance comparison: mAP and AUC-ROC for multi-label sentiment classification



(a) Style-wise mAP Comparison

(b) User study summary

(c) Ablation study

Fig. 8. mAP performance comparison: VQ-Emo vs. baseline across six artistic styles

VQ-Emo achieved a mAP of 0.712 for the Realism style, representing the optimal value across all styles. This indicates that even with style adaptation, the visual-emotional mapping relationship in realistic works remains more readily learnable by the model. Conversely, abstract styles exhibit the lowest mAP across all categories, despite style-adaptation optimization. This highlights the heightened subjectivity, ambiguity, and cultural-context dependence inherent in abstract art's emotional expression, presenting a key area for future breakthroughs in research.

6. Limitations and Future Work

6.1. Limitations

The VAWE-Art dataset constructed in this study comprises 5,000 AI-generated images spanning 10 mainstream artistic styles. However, due to annotation costs, niche styles such as Surrealism and Minimalism have only 300 samples per

category. This insufficient sample size hinders adequate modeling of visual-emotional mapping patterns for such styles. This is one core reason for the significantly lower emotional prediction performance observed in styles such as abstract expressionism compared to realistic styles in our experiments. Furthermore, the dataset primarily covers mainstream commercial generative models such as mid-journey and stable-diffusion, with insufficient coverage of works generated by niche open-source models or personalized LoRA models. The model's generalizability regarding extreme generation defects and niche custom styles remains to be validated. Furthermore, although emotional annotations passed Fleiss' Kappa consistency test, unavoidable subjective biases persist, as art professionals and amateur enthusiasts exhibit significant discrepancies in their annotations of the emotions in abstract works. Consistency remains lower for low-frequency emotional categories compared to high-frequency ones, and boundaries for compound emotions remain ambiguous.

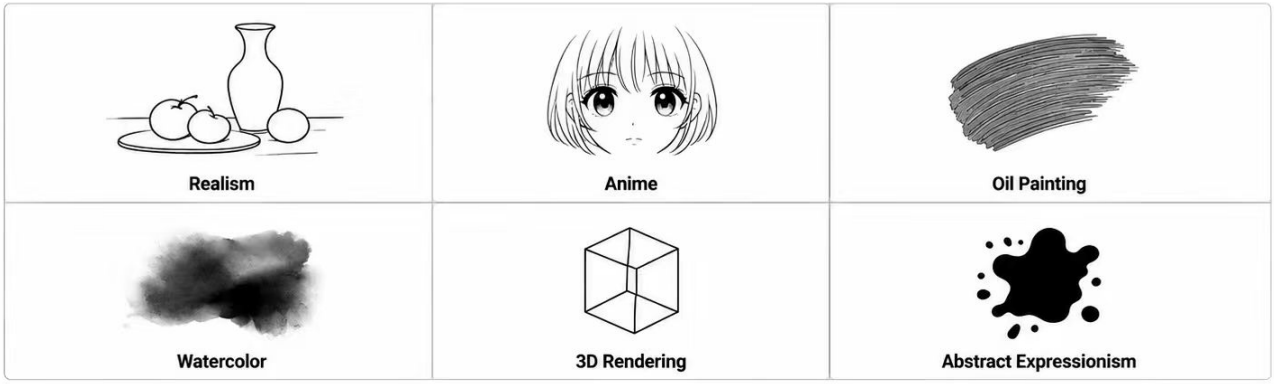


Fig. 9. Representative images of six artistic styles used in the style-specific analysis

VQ-Emo achieved a mAP of 0.712 for the Realism style, representing the optimal value across all styles. This indicates that even with style adaptation, the visual-emotional mapping relationship in realistic works remains more readily learnable by the model. Conversely, abstract styles exhibit the lowest mAP across all categories, despite style-adaptation optimization. This highlights the heightened subjectivity, ambiguity, and cultural-context dependence inherent in abstract art's emotional expression, presenting a key area for future breakthroughs in research.

6. Limitations and Future Work

6.1. Limitations

The VAWE-Art dataset constructed in this study comprises 5,000 AI-generated images spanning 10 mainstream artistic styles. However, due to annotation costs, niche styles such as Surrealism and Minimalism have only 300 samples per category. This insufficient sample size hinders adequate modeling of visual-emotional mapping patterns for such styles. This is one core reason for the significantly lower emotional prediction performance observed in styles such as abstract expressionism compared to realistic styles in our experiments. Furthermore, the dataset primarily covers mainstream commercial generative models such as midjourney and stable diffusion, with insufficient coverage of works generated by niche open-source models or personalized LoRA models. The model's generalizability regarding extreme generation defects and niche custom styles remains to be validated. Furthermore, although emotional annotations passed Fleiss' Kappa consistency test, unavoidable subjective biases persist: art professionals and amateur enthusiasts exhibit significant discrepancies in their annotations of the emotions in abstract works. Consistency remains lower for low-frequency emotional categories compared to high-frequency ones, and boundaries for compound emotions remain ambiguous.

The current VQ-Emo framework performs quality and emotional assessments solely on visual features extracted from images, without considering the artwork's art-historical context or stylistic background. Works sharing the same visual style may belong to different artistic movements or creative contexts, resulting in fundamentally distinct emotional connotations and aesthetic values. The model is unable to distinguish these context-driven emotional differences. Furthermore, the model's affective annotation and training rely solely on the aesthetic perceptions of domestic annotators, neglecting cross-cultural variations: audiences from different cultural backgrounds exhibit significant disparities in emotional responses to identical visual elements (e.g., Eastern audiences perceive Chinese ink paintings as evoking ‘Zen-like tranquility’ in stark contrast to Western interpretations). The current model's universality is constrained by a monocultural annotation system, rendering it unsuitable for cross-cultural aesthetic evaluation scenarios. To further clarify the limitations of the current VAWE-Art dataset, we acknowledge that it remains of moderate scale and exhibits structural imbalances. Although it includes 5,000 images across ten artistic styles, several sentiment categories and niche visual forms remain underrepresented. This imbalance restricts the model’s ability to fully capture fine-grained emotion–style relationships and limits generalizability to rare stylistic expressions. Additionally, while the dataset incorporates multiple mainstream generative models, its coverage of niche open-source and personalized LoRA models remains insufficient, limiting its scalability across broader generative ecosystems. These limitations indicate the need for future dataset expansion in both volume and diversity, including the incorporation of more styles, cultures, and generative pathways, ultimately enhancing the robustness, universality, and practical applicability of emotion-aware image evaluation.

6.2. Future Work

To address the current dataset limitations, subsequent efforts will expand the VAWE-Art dataset to over 20,000 samples, comprehensively covering niche art styles such as avant-garde art, traditional Chinese-style illustrations, and pixel art.

Additionally, works generated by more open-source and customized models will be incorporated to enhance the dataset's coverage of stylistic and generative pathways. Regarding annotation systems, cross-cultural annotators from at least six countries and regions, including China, the US, Europe, and Japan, will be engaged to establish a cross-cultural affective annotation framework. Concurrently, annotation workflows will be optimized by involving art history specialists in calibration to improve consistency in annotating low-frequency composite emotions. Supplementary metadata annotations, such as generation prompts, artistic movements, and creative contexts, will be introduced to provide data support for the model's multimodal expansion and contextual modeling. Modeling art-historical contexts and cross-cultural aesthetic differences. Subsequent work will construct an art-movement knowledge graph that integrates semantic information, such as art-historical backgrounds, stylistic characteristics, and creative contexts, into the VQ-Emo framework. A new context-encoding module will enable joint modeling of 'visual features + artistic context,' addressing emotional discrepancies arising from identical visual features across different contexts. Concurrently, by leveraging cross-cultural annotated datasets, we will develop a culturally sensitive emotion-prediction branch. This will model the mapping differences between visual elements and emotional responses across cultural contexts, enabling personalized emotion evaluation tailored to diverse audiences and expanding the model's cross-cultural universality.

7. Conclusion

This paper addresses the disconnect between visual quality and emotional impact assessment in AI-generated artworks, as well as the lack of a unified computational framework. It proposes a dual-modal, multi-dimensional, and interpretable joint evaluation method for the VQ-Emo framework. Through multi-branch collaborative modeling, style-aware adaptation, and multi-task joint optimization, this framework achieves end-to-end prediction of both visual quality scores and 20-dimensional emotional vectors for AI artworks. To support model training and evaluation, this study constructs the VAWE-Art dataset, comprising 5,000 AI-generated images spanning 10 mainstream artistic styles. It pioneers the provision of fine-grained visual quality scores and 20-dimensional emotional parallel annotations based on the VAWE Emotion Wheel, thereby filling a critical gap in the data resource for this field. Experimental results demonstrate that VQ-Emo achieves SRCC and PLCC scores of 0.912 and 0.925, respectively, in visual quality assessment, significantly outperforming traditional reference-free IQA methods and existing AIGC-specific evaluation approaches. In multi-label affective classification, it achieves a mAP of 0.678 and an AUC-ROC of 0.892, comprehensively surpassing general affective models and multimodal baselines. Ablation studies and style-specific analyses further validate the effectiveness of core designs, including the style-aware adaptation module, label dependency modeling, and consistency constraints, revealing quantifiable mapping patterns between visual features and emotional responses across diverse artistic styles. Moreover, through attention visualization and feature attribution analysis, the model provides interpretable visual evidence pinpointing key image regions that drive emotional responses. Whilst VQ-Emo demonstrates strong performance across mainstream styles, limitations persist in niche artistic styles, cross-cultural affective cognition, and art-historical context modeling. Future work will focus on dataset expansion, incorporating cross-cultural annotations, and integrating artistic context modeling to enhance the model's generalization and cultural adaptability. Overall, VQ-Emo establishes a unified computational framework for objectively evaluating AI-generated artworks while laying the methodological groundwork for emotion-controllable generative art systems. All code and data have been open-sourced to foster ongoing advancement in this field.

8. Managerial Implications

This study provides practical insights for managers in creative industries, digital design firms, and AI art platforms. Managers can incorporate the VQ-Emo framework into their decision-making processes in several ways. First, quality-control managers can use the visual-quality scores to screen and rank large volumes of AI-generated artworks before publication, reducing manual evaluation costs. Second, creative directors and product managers can use emotion predictions to match artworks with targeted audience groups or specific marketing goals, improving aesthetic alignment and emotional resonance. Third, platform operators can integrate the model into recommendation systems to ensure that displayed artworks align with users' emotional preferences. Finally, brand strategists may leverage the emotion-quality mapping to guide briefing instructions for AI-generated content, enabling more consistent emotional styles across campaigns. Beyond the empirical contributions, this research opens several long-term, thought-provoking questions. As AI systems increasingly quantify aesthetic and emotional properties, the definition of artistic value may gradually shift toward algorithmically measurable dimensions. Additionally, the model highlights potential cross-cultural divergences in emotional perception, raising the question of whether future AI art systems should learn culturally adaptive emotional standards. Finally, the expanding role of emotion-aware generative systems may reshape the relationship between human creators and AI tools, prompting reflection on authorship, creativity, and the evolution of artistic norms in the age of computational aesthetics.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author used Grammarly solely for language editing, grammar correction, and readability improvement. No AI tools were used to generate scientific content, conduct data analysis, interpret results, or draw conclusions. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

Reference

Cheng, R. (2021). *Sentiment analysis of automotive review texts based on multi-label classification* (Doctoral dissertation, Donghua University).

Chow, L. S. and Rajagopal, H. (2017). Modified-BRISQUE as no-reference image quality assessment for structural MR images. *Magnetic Resonance Imaging*, 43, 74–81.

Khan, F. F., Kim, D., Jha, D., Mohamed, Y., Chang, H. H., Elgammal, A., Elliott L., Elhoseiny M. (2024). AI art neural constellation: Revealing the collective and contrastive state of AI-generated and human art. *IEEE*.

Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., and Lin, W. (2023). AGIQA-3K: An open database for AI-generated image quality assessment. *arXiv preprint arXiv:2306.04717*.

Li, F. (2022). *Research and application of image data augmentation techniques based on generative adversarial networks* (Doctoral dissertation, Zhejiang University).

Yang, J. (2025). The aesthetic representation of AI-generated images and their "irrelevance." *Studies in Art*, (3).

Yu, P. and Zhang, Y. (2024). From feature recognition to image generation: Miao ethnic costume design based on the AIGC paradigm. *Silk*, (3), 61.

Yuan, J., Cao, X., Cao, L., Che, J., Wang, Q., and Ren, W. (2026). Text-image encoder-based contrastive regression for AI-generated image quality assessment. In *Proceedings of the International Conference on Artificial Neural Networks*. Cham: Springer.

Zhao, S. C. (2016). *Computational and applied research on image affect perception* (Doctoral dissertation, Harbin Institute of Technology).

Zhao, X., Yuan, G., and Dong, L. (2025). Detection of various solar radio bursts based on stable diffusion and self-supervised pretraining. *Solar Physics*, 300(12), 171.

Zheng, J. (2016). *Research on multi-focus image fusion methods* (Doctoral dissertation, Chongqing University).

Gang Hengju graduated from Peking University with a specialization in Digital Art (2007). He is currently employed in the Department of Animation and Esports at the Shanghai College of Publishing and Printing, where his areas of expertise include digital art effects, graphic design and AI image generation. He also serves as a director of the Shanghai Illuminating Engineering Society and as executive publisher of the journal Lighting Design. He has previously acted as a reviewer for the journal and has published numerous articles in internationally renowned peer-reviewed journals and conference proceedings. His areas of interest include digital art, lighting design, and image processing.