Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.5, No.7, 2025

Vision-Language Model-Based Emotional Understanding and Interpretable Feedback for Art Therapy Support

Jui Yun Su[1*]

[1]School of Visual Arts, New York, USA

**Corresponding author: juiyunsu007@gmail.com*

Abstract: Artworks can evoke complex, subjective, and heterogeneous emotional responses among viewers, whereas conventional approaches often reduce such responses to a single emotion label and therefore fail to jointly preserve affective diversity, visually grounded evidence, and linguistic interpretation. This paper proposes a large vision-language framework for artwork-evoked emotion understanding and interpretable feedback generation, with potential relevance to professionally guided reflection and discussion in art-therapy-oriented settings. Given an artwork image and an affect-oriented instruction, visual and textual representations are jointly encoded, and bidirectional cross-modal attention is introduced to strengthen the correspondence between regional visual content and affective semantics. During emotion inference, an affective prototype memory is employed to estimate multi-category emotion distributions, while uncertainty estimation is incorporated to avoid overly confident interpretations of ambiguous artistic expressions. To improve interpretability, an evidence-sensitive grounding mechanism identifies and aggregates color, subject, texture, and compositional regions associated with the predicted affective responses, producing a visually grounded emotion representation. An explanation decoder subsequently generates natural-language feedback conditioned on the predicted emotion distribution, regional visual evidence, and confidence information, while semantic consistency constraints maintain agreement among affective predictions, grounded regions, and explanatory text. Comparative experiments on the ArtEmis dataset demonstrate strong overall performance in terms of BLEU-1, BLEU-4, METEOR, and ROUGE-L. The results indicate that the proposed framework improves lexical correspondence, content coverage, and linguistic consistency in artwork emotion explanations, providing a reliable technical foundation for organizing artwork-evoked emotional responses and visual evidence in art-therapy-oriented auxiliary analysis.

Keywords: Art therapy support; visual language big data model; emotion distribution prediction; interpretable feedback.

1 Introduction

With the continuous changes in social environment, academic pressure, and interpersonal relationships, emotional distress, expression difficulties, and psychological adjustment problems among students and patients have gradually received widespread attention[1]. Traditional psychological support typically relies on interviews, scales, and professional observation; however, some individuals, due to limitations in age, language ability, cognitive level, or willingness to express themselves, find it difficult to accurately describe their inner feelings through direct language. Art therapy, through non-verbal forms such as painting, color, composition, and symbols, provides a gentler pathway for emotional expression and psychological communication, reducing the pressure of direct communication and providing rich visual information for understanding an individual's emotional experience. Therefore, how to identify potential emotional cues from artworks and transform them into valuable supplementary information has become an important research direction in the field of art therapy support[2].

Emotional expression in artworks is highly subjective, implicit, and ambiguous. Visual factors such as color and brightness, line shapes, spatial layout, character relationships, and scene semantics often work together and are influenced by the creator's experience, cultural background, and expressive habits. Traditional image recognition methods typically focus on object categories, local textures, or single emotion labels, making it difficult to fully understand the complex emotional semantics within artworks and to generate interpretations that align with human cognitive logic. In art therapy support scenarios, simply providing emotion categories is insufficient to meet professionals' needs for further understanding of judgment criteria, visual cues, and emotional context. Therefore, it is necessary to construct intelligent analysis methods that integrate emotion perception, semantic understanding, and natural language expression capabilities.

The development of large-scale visual language models provides a new technological foundation for understanding the emotions in art images. These models can jointly process visual content and natural language information, comprehensively describing potential emotions and their causes while identifying image subjects, scene relationships, and visual attributes[3]. Compared to traditional classification models, large-scale visual language models can link emotion judgments with color features, composition, character states, and symbolic content, expanding the analysis results from single labels to semantically structured emotional interpretations. Introducing large-scale visual language models into art therapy support helps improve the comprehensibility of emotional information in artworks and provides professionals with more intuitive, continuous, and evidence-based support feedback.

Research on emotion understanding for art therapy support has significant theoretical and practical value. From a theoretical perspective, this direction can promote the interdisciplinary integration of computer vision, natural language processing, affective computing, and human-computer interaction, and facilitate collaborative modeling between visual features, emotional semantics, and interpretive text. From an application perspective, interpretable emotional feedback can provide supplementary references for school psychological support, patient emotional communication, and expressive art activities, helping professionals more efficiently organize and understand emotional cues in artworks.

Simultaneously, related technologies should adhere to their supportive role, avoiding directly equating model outputs with psychological diagnostic results[4]. This will allow for the expansion of the practical application space of artificial intelligence in art therapy support scenarios while ensuring safety, interpretability, and professional boundaries.

In this study, emotional understanding refers specifically to the recognition and explanation of affective responses evoked by an artwork in viewers. It does not aim to infer the creator's internal psychological state, diagnose mental health conditions, or determine the therapeutic needs of an individual. Accordingly, the proposed framework is positioned as an auxiliary tool for organizing artwork-related emotional responses and visual evidence, which may support professionally guided reflection and discussion in art-therapy-oriented settings.

2 BackGround

Art therapy, using visual creation as its primary medium, conveys feelings and experiences that are difficult for individuals to express directly through forms such as painting, collage, color selection, and graphic organization[5]. Compared to support methods that rely solely on verbal communication, art creation can present emotional states, focus of attention, and relational experiences with lower expressive pressure, thus being widely used in scenarios such as school counseling, rehabilitation support, and emotional communication. However, the information contained in artworks is usually dispersed across multiple visual layers; the same color, symbol, or composition may have different meanings in different contexts. Professional analysis requires not only attention to the image content but also a comprehensive understanding that considers the emotional context, mode of expression, and individual differences. This highly experience-dependent analytical process increases the difficulty of manual processing and long-term tracking, creating a real need for standardized auxiliary tools[6].

In recent years, affective computing has gradually expanded from facial expression, speech, and text analysis to art image understanding, and the research focus has shifted from single-category judgment to emotional semantic modeling, visual evidence extraction, and language interpretation generation[7]. Visual language big data models, by unifying the representation of images and text, can simultaneously handle subject content, color relationships, spatial structure, and semantic description, providing a new path for organizing emotional information in artworks. Its core value lies not only in identifying potential emotional tendencies but also in transforming model judgments into readable visual evidence and semantic feedback, thereby enhancing the transparency of the analysis results. Building art therapy support systems based on such models is expected to create a complete information chain from image perception and emotion understanding to interpretive feedback and provide a foundation for subsequent human-machine collaborative decision-making, individual change recording, and support strategy development.

Large language model research has also established useful principles for preserving semantic intent under task-specific adaptation. Task-aware regularized continual fine-tuning limits representational drift during iterative adaptation [8], while masked language modeling improves the robustness of semantic matching under incomplete or noisy wording [9]. Instruction-driven cross-attention further shows how task queries can steer classification-relevant feature fusion [10], and semantic alignment mechanisms

provide a complementary basis for keeping user-facing feedback responsive to an intended interpretive context [11]. These considerations support the use of affect-oriented instructions and consistency constraints in artwork-centered emotional feedback.

3 Methodology

Rather than reducing an artwork to a single categorical judgment, the proposed method formulates emotional understanding as a joint inference problem involving subjective affect distribution, visually grounded evidence, and natural-language feedback. Given an artwork image $\boldsymbol{I}$ and an affect-oriented instruction $\boldsymbol{p}$, a vision-language backbone first establishes a shared representational space in which visual structure and linguistic intent can interact without discarding the ambiguity inherent in artistic expression. To preserve both local details and global composition, the image encoder transforms the artwork into a sequence of region-sensitive visual tokens, while the text encoder converts the instruction into semantic query tokens that specify the required emotional interpretation:

$$\boldsymbol{V} = \mathcal{F}_v(\boldsymbol{I}), \qquad \boldsymbol{T} = \mathcal{F}_t(\boldsymbol{p}) \tag{1}$$

By retaining the spatial arrangement of $\boldsymbol{V} \in \mathbb{R}^{N\times d}$ and the contextual semantics of $\boldsymbol{T} \in \mathbb{R}^{M\times d}$, where N and M denote the numbers of visual and textual tokens and d represents the shared embedding dimension, the model can associate emotional language with specific compositional regions instead of relying only on global appearance. Beyond conventional feature concatenation, bidirectional cross-modal interaction is introduced to refine both modalities under a common affective context:

$$\widetilde{\boldsymbol{V}} = MHA(\boldsymbol{V}, \boldsymbol{T}, \boldsymbol{T}) + \boldsymbol{V}, \qquad \widetilde{\boldsymbol{T}} = MHA(\boldsymbol{T}, \boldsymbol{V}, \boldsymbol{V}) + \boldsymbol{T} \tag{2}$$

Such reciprocal refinement allows linguistic intent to emphasize emotionally relevant image regions while permitting visual evidence to constrain the semantic scope of the textual representation. The overall architecture of the model is shown in Figure 1.

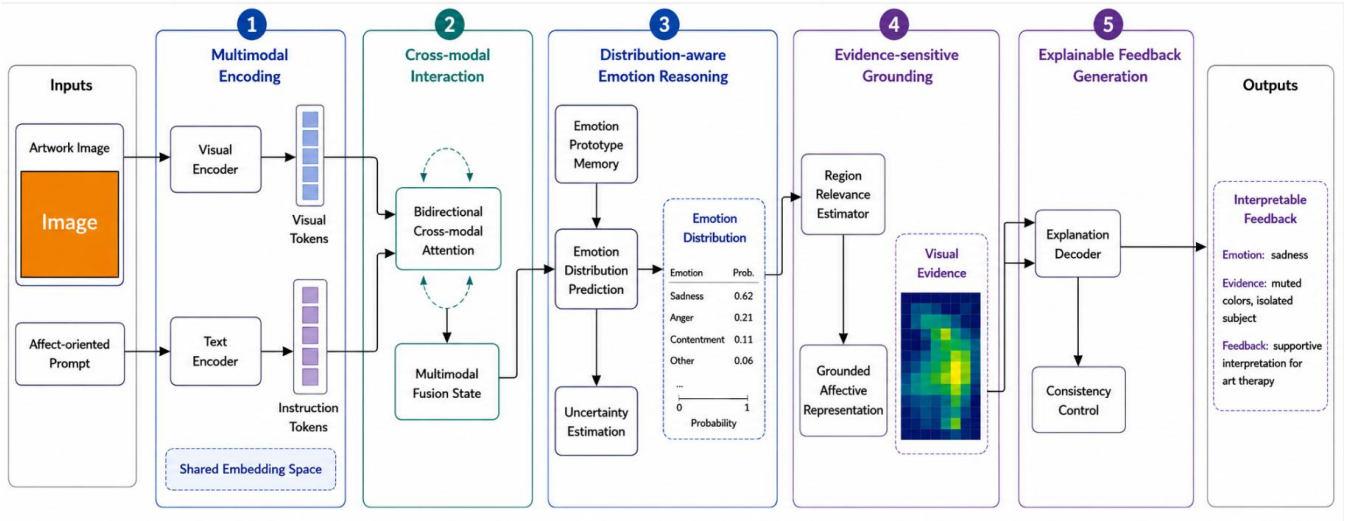


Figure 1. Overall Model Architecture

The cross-modal representation is also designed to remain stable when visual cues are heterogeneous or partially ambiguous. Consistency-aware unified modeling [12] and self-supervised multi-scale learning for resource-state prediction [13] motivate the integration of complementary local

and global signals. Multi-scale convolution with adaptive attention [14] and graph-based anomaly modeling in multi-tenant environments [15] further demonstrate the value of selectively emphasizing salient dependencies rather than treating all inputs uniformly. Lightweight detection strategies for distributed edge-cloud settings [16] likewise support a compact, evidence-centered design suitable for efficient multimodal inference.

A distribution-aware affective representation is then constructed to avoid forcing intrinsically subjective responses into a rigid one-hot label. Because multiple viewers may report different emotions for the same artwork, the supervision signal is expressed as an empirical probability distribution whose elements reflect the relative frequency of each affective category:

$$q_k = \frac{n_k + \epsilon}{\sum_{j=1}^{K} (n_j + \epsilon)} \tag{3}$$

For the kth emotion category, n_k denotes the number of corresponding annotations, K is the affected vocabulary size, and ϵ provides numerical smoothing while preserving minority emotional responses. Grounded on the fused visual and textual states, an affective prototype space estimates the predicted distribution through similarity-based matching rather than an isolated classification layer:

$$\widehat{p}_k = \frac{\exp\left(\tau^{-1}\mathrm{sim}\left(\mathbf{z}, \mathbf{e}_k\right)\right)}{\sum_{j=1}^{K} \exp\left(\tau^{-1}\mathrm{sim}\left(\mathbf{z}, \mathbf{e}_j\right)\right)} \tag{4}$$

Instead of treating $\mathbf{e}_k$ as a purely numerical classifier weight, each prototype represents the semantic center of one emotion, whereas $\mathbf{z}$ summarizes the multimodal state and τ controls the sharpness of the resulting probability distribution.

An evidence-sensitive grounding mechanism further restricts emotional reasoning to visually supportable content. To identify the regions that materially contribute to the inferred effect, the multimodal state generates a normalized relevance vector over all visual tokens:

$$\alpha_i = \frac{exp\left(\boldsymbol{w}_g^{\top} tanh\left(\boldsymbol{W}_v \tilde{\boldsymbol{v}}_i + \boldsymbol{W}_z \boldsymbol{z}\right)\right)}{\sum_{r=1}^{N} exp\left(\boldsymbol{w}_g^{\top} tanh\left(\boldsymbol{W}_v \tilde{\boldsymbol{v}}_r + \boldsymbol{W}_z \boldsymbol{z}\right)\right)} \tag{5}$$

Using α_i as the contribution assigned to region i, the grounded affect vector aggregates relevant color, subject, texture, and compositional cues while suppressing visually unrelated content:

$$\boldsymbol{g} = \sum_{i=1}^{N} \alpha_i \tilde{\boldsymbol{v}}_i \tag{6}$$

Instead of producing unrestricted free-form text, the language decoder receives $\boldsymbol{g}$ together with the predicted affect representation so that each generated statement remains connected to observable visual evidence:

$$P(\mathbf{y}|\mathbf{I}, \mathbf{p}) = \prod_{t=1}^{L} P\left(y_t | y_{<t}, \mathbf{g}, \sum_{k=1}^{K} \widehat{p}_k \mathbf{e}_k\right) \tag{7}$$

This conditioning strategy gives the feedback a dual role by communicating the likely emotional interpretation and clarifying which visual characteristics support that interpretation.

Uncertainty is incorporated to prevent ambiguous artworks from receiving overly assertive feedback. During optimization, distribution matching, explanation generation, visual grounding,

semantic consistency, and confidence calibration are jointly balanced so that interpretability is treated as part of the learning objective rather than as a post hoc visualization procedure:

The reliability component also benefits from treating confidence as a context-sensitive property of the underlying signals. Attention-enhanced traffic prediction for load balancing [17] and topology-aware graph neural diagnosis [18] illustrate how structured dependencies can sharpen state estimation, while reinforcement learning under energy constraints [19] highlights the need to balance utility with bounded operational risk. Multi-scale Transformer forecasting of CPU utilization [20], collaborative discrimination of reconstruction residuals and latent shifts [21], and contrastive masked autoencoding for distributed task representations [22] collectively motivate calibration that responds to variability, distributional uncertainty, and the strength of available evidence.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{KL}} + \lambda_2 \mathcal{L}_{\mathrm{gen}} + \lambda_3 \mathcal{L}_{\mathrm{ground}} + \lambda_4 \mathcal{L}_{\mathrm{cons}} + \lambda_5 \mathcal{L}_{\mathrm{cal}} \tag{8}$$

Here, $\mathcal{L}_{\mathrm{KL}}$ aligns the predicted and observed effect distributions, $\mathcal{L}_{\mathrm{gen}}$ preserves linguistic fidelity, $\mathcal{L}_{\mathrm{ground}}$ encourages evidence concentration on relevant regions, $\mathcal{L}_{\mathrm{cons}}$ maintains agreement between the stated emotion and the generated rationale, and $\mathcal{L}_{\mathrm{cal}}$ discourages unjustified confidence through reliability-aware regularization. Overall, the resulting formulation supports art-therapy-oriented analysis by combining probabilistic emotional interpretation, region-level visual evidence, and cautious explanatory feedback within a unified vision-language framework.

4 Experimental Results and Analysis

4.1 Dataset

The ArtEmis dataset was adopted to support the study of emotion understanding and explanation generation for visual artworks. It contains approximately 80,000 artwork images associated with nearly 455,000 human annotations, and each annotation includes an emotion category together with a natural language explanation describing the visual or semantic factors that evoke the corresponding affective response. The emotional vocabulary covers amusement, awe, contentment, excitement, anger, disgust, fear, sadness, and other responses, thereby providing both categorical supervision and linguistically rich affective descriptions. Multiple annotations are available for many artworks, which preserve the subjective diversity of human emotional perception and enable the construction of empirical emotion distributions rather than restricting each image to a single deterministic label. The combination of artwork images, affective categories, and explanatory text makes the dataset suitable for jointly learning visual representations, emotion semantics, region-related evidence, and interpretable feedback within a unified vision-language framework.

4.2 Experimental setup

All experiments were conducted under a unified hardware and software environment to ensure reproducibility and consistent optimization across the emotion distribution prediction, visual evidence grounding, uncertainty estimation, and explanation generation tasks. The 32-billion-parameter Qwen3-VL-32B-Instruct model was adopted as the vision-language backbone and was adapted using 4-bit quantized low-rank fine-tuning, which reduced memory consumption while preserving the multimodal reasoning capability required for artwork understanding. Training was performed on a single NVIDIA

H100 GPU with 80 GB of memory using bfloat16 computation, gradient accumulation, gradient checkpointing, and a cosine learning-rate schedule with linear warm-up. Artwork images were resized to 448 × 448 pixels, and the text sequence lengths were constrained to maintain a balance between explanatory completeness and computational efficiency. The complete implementation environment and principal hyperparameter configuration are summarized in Table 1.

Table 1. Experimental Environment and Hyperparameter Settings

Category	Configuration Item	Setting
Hardware	GPU	1 × NVIDIA H100 80 GB
Hardware	CPU	Multi-core server processor
Hardware	System memory	256 GB
Software	Operating system	Ubuntu 22.04 LTS
Software	Python	3.10
Software	Deep learning framework	PyTorch 2.5.1
Software	CUDA	12.4
Software	Transformers	4.57.1
Software	Parameter-efficient tuning	PEFT 0.17.1
Model	Vision-language backbone	Qwen3-VL-32B-Instruct
Model	Fine-tuning strategy	4-bit QLoRA
Model	Computation precision	bfloat16
Model	Image resolution	448 × 448
Model	Maximum input length	2048 tokens
Model	Maximum output length	256 tokens
Training	Training epochs	3
Training	Optimizer	AdamW
Training	Initial learning rate	1×10^{-4}
Training	Weight decay	0.01
Training	Learning-rate schedule	Cosine decay
Training	Warm-up ratio	0.03
Training	Per-device batch size	1
Training	Gradient accumulation steps	16
Training	Effective batch size	16
Training	Maximum gradient norm	1.0
Training	LoRA rank	64
Training	LoRA scaling factor	128
Training	LoRA dropout	0.05
Training	Gradient checkpointing	Enabled
Training	Random seed	42
Inference	Decoding temperature	0.2
Inference	Top-p sampling	0.9

4.3 Experimental Results and Analysis

To evaluate the overall performance of different methods in visual artwork emotion understanding and interpretable explanation generation, representative approaches covering visual emotion analysis, artwork affect modeling, vision-language explanation generation, and affective image captioning were selected as comparison methods. To ensure a fair and consistent comparison, all methods were reimplemented and retrained under the same experimental protocol, using identical dataset partitions, image preprocessing procedures, training settings, and evaluation scripts. The generated explanations were assessed from complementary perspectives, including local lexical correspondence, long-range

phrase consistency, semantic alignment, and content coverage. Accordingly, BLEU-1 (B1), BLEU-4 (B4), METEOR (M), and ROUGE-L (R) were employed as the quantitative evaluation metrics. The comparative results obtained after unified retraining are reported in Table 2.

Table 2. Comparison of representative methods on ArtEmis-related benchmarks

Method	B1	B4	M	R
Achlioptas et al.[23]	53.6	8.7	14.2	29.7
Borth et al.[24]	39.6	1.4	8.8	20.2
Cornia et al.[25]	50.7	9.5	14.0	28.0
Nguyen et al.[26]	70.1	11.3	16.8	33.3
Mokady et al.[27]	47.8	9.3	16.0	25.6
Sammani et al.[28]	53.6	8.4	13.6	30.0
Mohamed et al.[29]	62.8	13.7	16.5	33.9
Ours	73.9	16.7	18.9	35.4

B1 denotes BLEU-1, which measures unigram-level lexical correspondence between the generated explanation and the reference text, while B4 represents BLEU-4 and evaluates the consistency of longer four-gram sequences. M refers to METEOR, which considers word alignment, stemming, synonym matching, and recall-oriented semantic correspondence. R denotes ROUGE-L, which assesses the similarity of sentence structure and content coverage through the longest common subsequence between generated and reference descriptions. Together, these four metrics provide a comprehensive evaluation of word-level accuracy, phrase-level fluency, semantic relevance, and overall information completeness.

The proposed method demonstrates strong performance across all four evaluation metrics, indicating that the generated emotional explanations maintain accurate lexical selection, coherent phrase organization, reliable semantic alignment, and sufficient content coverage. Its balanced results also show that the model can connect artwork content with affective semantics while producing explanations that remain consistent with human annotations. Such performance reflects the effectiveness of jointly modeling emotion distributions, visual evidence, and language generation, thereby supporting more informative and interpretable feedback for art therapy applications.

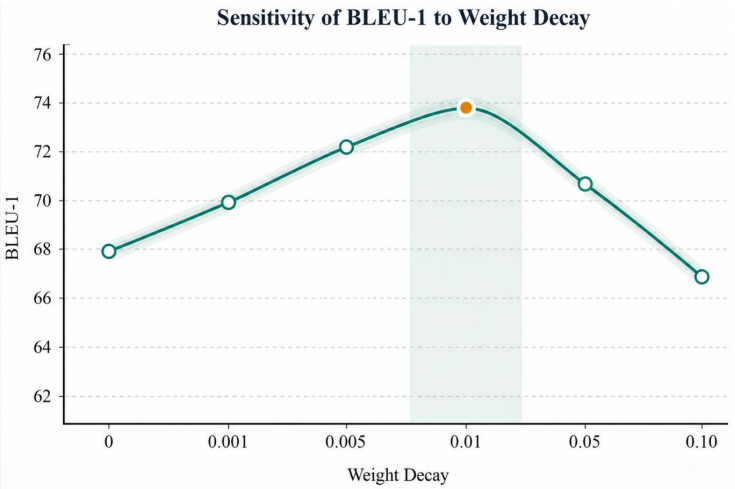


Figure 2. The effect of weight decay on the results of the B1 experiment

As shown in Figure 2, the weight decay analysis shows that the proposed method maintains strong BLEU-1 performance under moderate regularization, indicating that an appropriate constraint can improve lexical accuracy while preserving the expressive capacity of the vision language model. The selected configuration provides a favorable balance between parameter flexibility and model stability, allowing the generated emotional explanations to remain closely aligned with the reference descriptions. Excessively weak or strong regularization reduces this balance, whereas the adopted setting enables the proposed algorithm to produce more accurate and reliable affective language for artwork interpretation.

5 Conclusion

This paper addresses the need for emotional understanding and feedback in art therapy support scenarios. It constructs a unified analytical framework based on a large visual language model, integrating visual content, emotional semantics, regional evidence, and natural language interpretation from artworks into a single modeling process. This method moves beyond limiting artistic emotional understanding to single-category judgments. Instead, it preserves the potential subjective differences among observers regarding the same artwork through emotion distribution prediction and establishes a connection between image structure and emotional language through cross-modal interaction mechanisms. Building upon this, regional relevance modeling further extracts color, subject, composition, and spatial information closely related to emotion judgment, ensuring that the generated feedback corresponds to observable visual evidence. Uncertainty estimation and semantic consistency constraints reduce the risk of overjudgment in ambiguous artworks, fostering a more harmonious relationship between emotion recognition, visual evidence, and interpretive text. This provides intelligent analytical results with readability, traceability, and auxiliary value for art therapy applications.

Comparative experiments show that the proposed method demonstrates good comprehensive capabilities across evaluation dimensions such as word matching, long phrase organization, semantic relevance, and content coverage. This indicates that the collaborative design between emotion distribution modeling, visual evidence constraints, and interpretation generation effectively improves the quality of artistic emotional feedback. Compared to methods that focus solely on image content recognition or general description generation, this framework places greater emphasis on the emotional connotations within artworks and the underlying reasons for their formation. The output can simultaneously answer what emotions a work might evoke and which visual factors support this judgment. This analytical approach helps propel artificial intelligence from outcome-oriented emotion classification tools to process-transparent assisted understanding tools, providing a scalable technological foundation for school psychological support, patient emotional communication, rehabilitation interventions, expressive art activities, and digital psychological services. By consistently organizing and interpreting a large number of artworks, the system can also alleviate the workload of professionals in artwork screening, emotional cue summarization, and long-term record management, and promote a more prudent and effective collaborative relationship between artificial intelligence and art therapy practice.

Future research can further expand data sources across different ages, cultural backgrounds, art media, and application scenarios to enhance the model's adaptability to diverse artistic expressions and

reduce the limitations of a single data distribution on emotion understanding results. The dynamic creative process also deserves inclusion in future research. By combining brushstroke sequence, drawing duration, color switching, pause behavior, textual self-narration, and verbal expression, the analysis can be gradually expanded from static artwork analysis to joint modeling of continuous behavior and emotional changes. Meanwhile, research geared towards real-world applications needs to strengthen feedback and review mechanisms involving professionals, protect individual privacy, assess model bias, and control safety boundaries to avoid misusing supplementary interpretations as definitive psychological diagnostic conclusions. With the continuous development of multimodal foundational models, trustworthy artificial intelligence, and human-machine collaborative technologies, intelligent systems supporting art therapy are expected to provide more gentle, personalized, and understandable communication channels for groups with difficulties in emotional expression, while maintaining professional guidance and ethical constraints. This will also promote more practically valuable interdisciplinary research directions among art therapy, affective computing, and digital health services.

References

[1] J. Zhang, L. Zheng, M. Wang, et al., "Training a small emotional vision language model for visual art comprehension," in European Conference on Computer Vision, Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 397-413.

[2] J. Zhang, D. Guo, Z. Li, et al., "EmoSEM: Segment and explain emotion stimuli in visual art," in Proceedings of the 7th ACM International Conference on Multimedia in Asia, 2025, pp. 1-7.

[3] Y. Lin, J. Sun, Z. Q. Cheng, et al., "Why we feel: Breaking boundaries in emotional reasoning with multimodal large language models," in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 5196-5206.

[4] F. Man, X. Chen, H. Wang, et al., "KEVER2: Knowledge-enhanced visual emotion reasoning and retrieval," arXiv e-prints, p. arXiv:2505.24342, 2025.

[5] S. E. Lee, Y. Lee, and E. Park, "EmoVLM-KD: Fusing distilled expertise with vision-language models for visual emotion analysis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 5672-5681.

[6] D. Huang, Q. Li, C. Yan, et al., "Emotion-Qwen: A unified framework for emotion and vision understanding," arXiv preprint arXiv:2505.06685, 2025.

[7] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny and L. J. Guibas, "Artemis: Affective Language for Visual Art", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11569-11579, 2021.

[8] J. Hu, Y. Lyu, and J. Jiang, "Task-Aware Regularized Continual Fine-Tuning for Mitigating Catastrophic Forgetting in Large Language Models," 2025.

[9] Z. Ke, "Semantic User-Advertisement Alignment via Large Language Models for Personalized Recommendation," 2024.

[10] Y. Jiang and Y. L. Peng, "Instruction-Driven Language Model for Text Classification via Cross-Attention Fusion and Semantic Alignment," 2024.

[11] Y. Yang, M. Nagel, and A. Li, "Consistency-Aware Unified Modeling for Multi-Source System Anomaly Detection," 2025.

[12] X. Liu and K. Gong, "Predictive GPU Memory Modeling in Containerized Cloud Environments Using Self-Supervised Multi-Scale Learning," 2025.

[13] S. Pi, "Intelligent Backend Latency Forecasting with Multi-Scale Convolution and Adaptive Attention Mechanisms," 2024.

[14] S. Shawulieti, "Enhancing Cloud Security with Graph Neural Network-Driven Anomaly Detection in Multi-Tenant Environments," 2024.

[15] S. Xu, "Efficient Anomaly Detection in Distributed Edge-Cloud Systems Using Lightweight Modeling," 2024.

[16] Z. Qing and F. Huang, "Attention-Enhanced Network Traffic Prediction for Intelligent Load Balancing Optimization," Transactions on Computational and Scientific Methods, vol. 5, no. 8, 2025.

[17] F. Xue, K. Wu, and A. Abdulla, "Topology-Aware Graph Neural Networks for Distributed Node Fault Identification," 2025.

[18] J. Sun, "Reinforcement Learning-Based Task Scheduling Under Energy Budget Constraints in Edge-Cloud Systems," 2024.

[19] Z. Wang and S. Lin, "Multi-Scale Transformer for CPU Utilization Time Series Forecasting in Cloud Computing," 2025.

[20] Z. Bian and X. Su, "Enhancing Semantic Matching and Text Robustness through Masked Language Modeling," 2024.

[21] Z. Liu and D. Abdurehim, "Distributed System Anomaly Detection via Collaborative Discrimination of Reconstruction Residuals and Latent Distribution Shifts," 2025.

[22] J. Zhou, "Contrastive Masked Autoencoding for Self-Supervised Anomaly Representation Learning in Distributed Task Systems," 2024.

[23] P. Achlioptas, M. Ovsjanikov, K. Haydarov, et al., "ArtEmis: Affective language for visual art," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11569-11579.

[24] D. Borth, R. Ji, T. Chen, et al., "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 223-232.

[25] M. Cornia, M. Stefanini, L. Baraldi, et al., "Meshed-memory transformer for image captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578-10587.

[26] V. Q. Nguyen, M. Suganuma, and T. Okatani, "GRIT: Faster and better image captioning transformer using dual visual features," in European Conference on Computer Vision, Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 167-184.

[27] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP prefix for image captioning," arXiv preprint arXiv:2111.09734, 2021.

[28] F. Sammani, T. Mukherjee, and N. Deligiannis, "NLX-GPT: A model for natural language explanations in vision and vision-language tasks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8322-8332.

[29] Y. Mohamed, F. F. Khan, K. Haydarov, et al., "It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21263-21272.