AffectiveArt Challenge 2026: Fine-Grained Emotion Understanding and Generation in Artistic Images

A Proposal for the ACM Multimedia 2026 Grand Challenge

Abstract

Recent advances in AI-generated content (AIGC), driven by large-scale diffusion and multimodal foundation models, have significantly improved visual realism. However, current systems remain limited in modeling the fine-grained emotional expression and abstract visual language inherent in artistic imagery. Unlike photo-realistic generation, artistic creation requires coherent integration of style, brushwork, color composition, and affective intent. To address this gap, we propose the **EmoArt 2026 Challenge**, built upon the large-scale EmoArt dataset comprising 132,664 artworks across 56 artistic styles with structured affective annotations. The challenge consists of two complementary tracks: (1) *Emotion-Aware Artistic Image Generation*, focusing on emotionally and stylistically coherent synthesis; and (2) *Multidimensional Art Emotion Understanding*, targeting fine-grained affect recognition in artistic imagery. We introduce a rigorous evaluation protocol combining automatic metrics (including the proposed Attribute Alignment Score) and expert-based human aesthetic assessment. The hidden test set is securely maintained to ensure fair benchmarking and prevent data leakage. The challenge will be co-located with the *MUSE 2026 Workshop on Multimodal Understanding and Synthesis of Emotion in Art* at ACM Multimedia 2026, fostering close interaction between benchmark development and in-depth academic discussion. We envision EmoArt 2026 as a sustainable benchmark initiative that advances affective computing and emotion-aware artistic AI beyond the conference year.

1 Introduction and Motivation

1.1 Why Emotion and Art Lag Behind

The recent "LLM era" has dramatically reshaped the AI landscape. With the scaling of large language models and the emergence of strong instruction-following and tool-using agents, AI systems have made striking progress in **logical reasoning**, **mathematical problem solving**, and **code generation**. In many benchmark settings, models can now produce step-by-step solutions, verify intermediate results, and even self-correct via external tools.

Yet, these advances do not automatically translate into **affective intelligence**. Emotion understanding and artistic appreciation are rooted in subjective human experience, cultural context, and subtle perceptual cues. In fine art, meaning is often conveyed indirectly through low-level and mid-level visual decisions—e.g., brushstroke rhythm, tonal contrast, color harmony, spatial tension, and compositional balance—rather than through explicit objects alone. As a result, even when a model "knows" the word *sad*, it may not reliably connect sadness to the visual mechanisms that induce the feeling.

This gap is especially consequential in the age of generative models. Text-to-image systems can already synthesize high-fidelity content, but they frequently treat emotion as a superficial style token, producing outputs that are semantically correct yet affectively ambiguous or inconsistent. Bridging this gap requires benchmarks and datasets that connect **what** is depicted with **how** emotion is expressed visually, enabling models to learn controllable affective generation and interpretable affective understanding.

1.2 The Challenge of Affective Artistic Image Generation and Understanding

Recent advances in text-to-image models, such as Stable Diffusion 3.5 [15] and FLUX.1 [6], have dramatically improved the realism and controllability of visual generation. However, achieving *emotionally faithful artistic expression* remains a fundamental challenge. As illustrated in Fig. 1, current systems exhibit systematic limitations in both affective generation and understanding.

In artistic generation, emotional intent is encoded through subtle interactions among composition, color harmony, style, and symbolic cues. A "sad" painting, for example, is not merely an image of a crying subject, but may rely on descending compositional lines, low-key illumination, and a cool, desaturated palette. Existing generative models, largely trained on web-scale photographic data, tend to prioritize semantic correctness over affective coherence, often producing images that match textual objects but fail to capture the intended emotional style.

A parallel limitation arises in artistic understanding. Current vision–language models frequently produce surface-level captions that describe visible objects while neglecting affective and aesthetic reasoning. They struggle to infer how formal visual elements contribute to emotional atmosphere. Together, these issues reflect a deeper gap: the lack of joint modeling of *emotion, style, and semantics* in artistic multimedia. Addressing this gap requires new benchmarks and evaluation frameworks that explicitly target emotion-aware artistic reasoning.

1.3 Limitations of Existing Benchmarks

The progress of emotion-aware AIGC has been hindered by the scarcity of high-quality, fine-grained datasets. As illustrated in Table 1, existing resources suffer from critical limitations:

- **Insufficient Scale**: Pioneering datasets like ArtPhoto (806 images) [7] and Emotion6 (1,980 images) are too small to train data-hungry diffusion models.

ACM Multimedia '26, Rio de Janeiro, Brazil

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

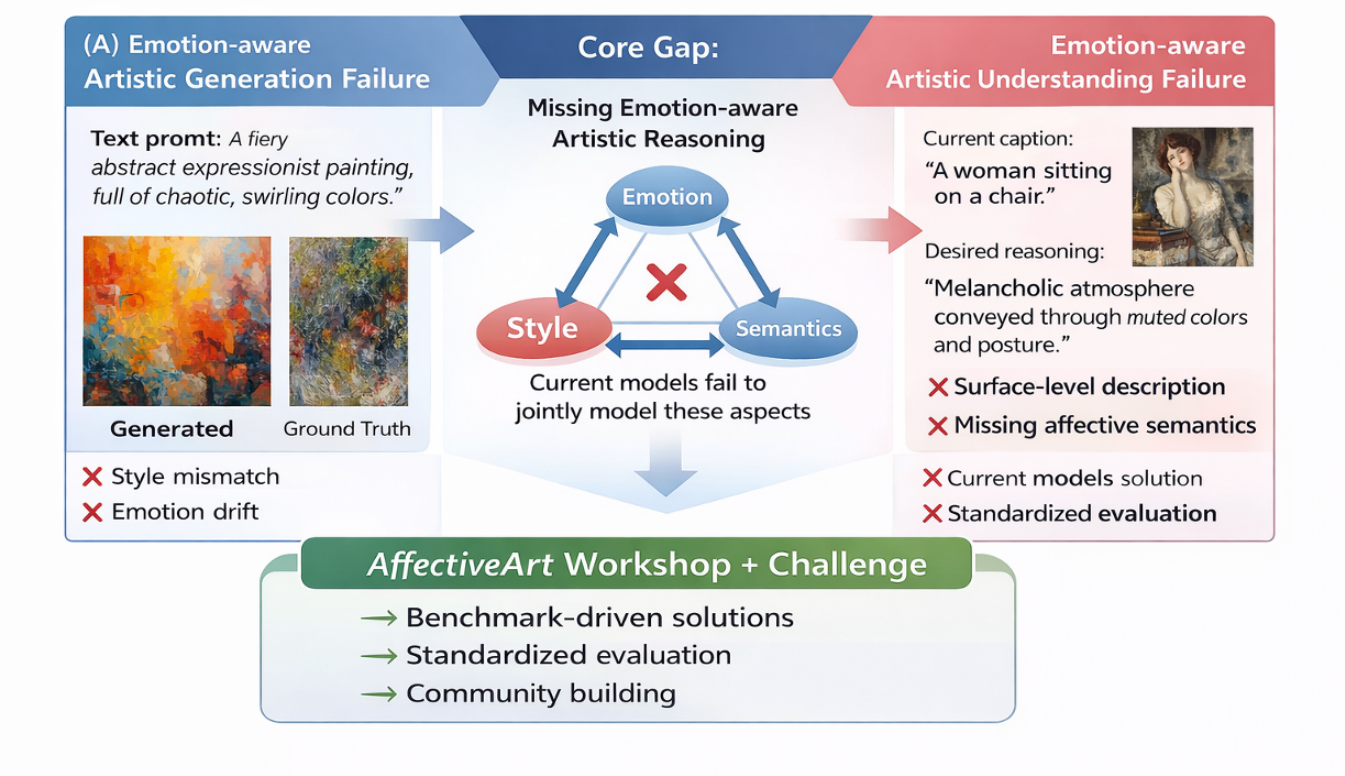

Figure 1: Current limitations of artistic AI in emotion-aware generation and understanding. Left: emotion-aware artistic generation frequently exhibits *style mismatch* and *emotion drift* with respect to the intended artistic expression. Right: artistic understanding often remains at surface-level captioning, lacking *affective semantics* and structured reasoning about composition and color relationships. Center: both limitations stem from a shared core gap—the absence of joint modeling of *emotion, style, and semantics*. The proposed AffectiveArt Grand Challenge + Workshop (MUSE 2026: Multimodal Understanding and Synthesis of Emotion in Art) aims to address this gap through benchmark-driven research, standardized evaluation protocols, and community building.

- **Domain Mismatch**: Datasets like EmoSet [16] focus on social media photography, which lacks the stylistic diversity and abstract nature of fine art.
- **Shallow Annotation**: While ArtEmis [1] provides affective captions, it lacks structured annotations for the visual attributes that *cause* the emotion. Without knowing why an image is perceived as "calm" (e.g., due to "smooth, blending brushstrokes"), models cannot learn to generate "calm" art effectively.

1.4 Our Contribution: The EmoArt Paradigm

The EmoArt 2026 Challenge seeks to redefine this landscape. We leverage the EmoArt dataset [19], which is distinct in its multi-dimensional approach. It creates a bridge between **Computer Vision**, **Art History**, and **Psychology**. By incorporating labels derived from art therapy (e.g., "Relieve Stress"), we also open new avenues for AI in mental health and well-being applications.

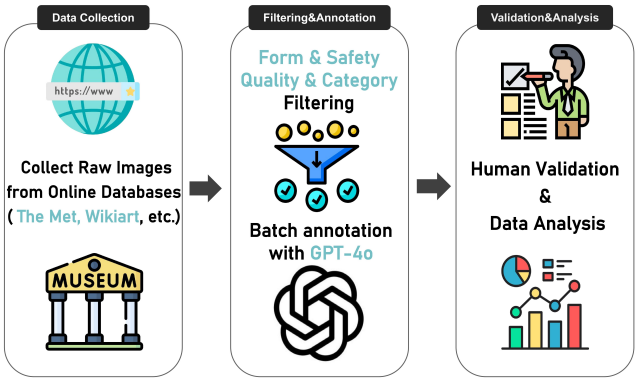

Figure 2: Construction pipeline of the EmoArt dataset.

2 The EmoArt Dataset

The foundation of this challenge is the EmoArt dataset, constructed through a meticulous pipeline ensuring legal compliance, cultural diversity, and annotation quality.

Table 1: Comparison of Emotion-related Datasets, R represents Recognition, G represents Generation.

Dataset	Image Type	Label Source	Tasks	Image	Category	Valence&Arousal	Attributes	Description
IAPSa[9]	Photo	Human	R	395	✓	✓	✗	✗
GAPED[3]	Photo	Human	R	730	✓	✓	✗	✗
ArtPhoto[7]	Art	Human	R	806	✓	✗	✗	✗
Emotion6[11]	Photo	Human	R	1980	✓	✗	✗	✗
FI[18]	Photo	Human	R	23308	✓	✗	✗	✗
WEBEmo[10]	Photo	Human	R	268K	✓	✗	✗	✗
Artemis[1]	Art	Human	G&R	80K	✓	✗	✗	✓
EmoSet[16]	Photo/Art	Human&LLM	G&R	3300K	✓	✗	✓	✗
FindingEmo[8]	Photo	Human	R	25K	✓	✓	✗	✗
EmoArt (Ours)	Art	Human&LLM	G&R	130K	✓	✓	✓	✓

2.1 Data Collection and Filtering

We aggregated over 200,000 raw images from authoritative public domain sources: WikiArt, The Metropolitan Museum of Art, The National Museum of Asian Art, Europeana, and the National Palace Museum.

To ensure the dataset is robust for machine learning tasks, we applied a four-stage filtering protocol:

(1) **Art Form Filtering**: We strictly retained only **paintings**. Other media such as sculptures, ceramics, and architecture were excluded to maintain consistency in 2D visual feature learning.
(2) **Content Safety**: A rigorous safety pipeline was implemented. We combined automated NSFW classification with manual review to remove explicit content, ensuring the dataset is suitable for academic and public use.
(3) **Quality Assurance**: Images with resolutions below 300×300 pixels, or those containing significant watermarks, borders, or compression artifacts, were discarded to prevent model degradation.
(4) **Category Balancing**: To avoid long-tail distribution issues common in art datasets, we filtered out artistic styles with fewer than 400 samples.

The final dataset contains **132,664 images** spanning **56 distinct styles**, grouped into seven thematic domains: *Classics*, *Modern Edge*, *East Spirit*, *Chromatic Soul*, *Dream Visions*, *Form & Flow*, and *Social Mirror*.

2.2 Hierarchical Annotation Pipeline

We employed a state-of-the-art annotation strategy leveraging GPT-4o as a multimodal aesthetic engine, followed by human verification. This allowed us to generate descriptions that are not just descriptive but "emotionally aware." The annotation system covers five dimensions:

(1) **Content Description**: Unlike brief COCO captions, our descriptions (Avg. 35.6 words) provide detailed narratives of the scene, integrating emotional cues.
(2) **Visual Attributes**: Based on art psychology, we provide structured descriptions for five key elements:
- *Brushwork*: Stroke thickness, rhythm, texture (e.g., "impasto," "delicate").
- *Composition*: Symmetry, balance, focal points (e.g., "triangular composition").
- *Color*: Hue palettes, saturation levels, harmony.
- *Line*: Curvature, continuity, intensity.
- *Light*: Contrast, direction, atmospheric effects (e.g., "chiaroscuro").

(3) **Valence and Arousal (VA)**: Continuous values based on Russell's Circumplex Model [14].
(4) **Dominant Emotion**: Classification into 12 representative categories selected to cover the VA space:
- *High Arousal / Positive*: Excited, Happy, Aroused.
- *High Arousal / Negative*: Alarmed, Annoyed, Frustrated.
- *Low Arousal / Negative*: Sad, Bored, Tired.
- *Low Arousal / Positive*: Calm, Content, Glad.

(5) **Therapeutic Potential**: Labels indicating the artwork's potential psychological benefit (e.g., "Relieve Stress"), inspired by professional art therapy practices.

2.3 Human Validation and Data Analysis

To validate the automated annotations, we conducted a study with 5,600 images. Ten trained annotators assessed the quality. The results demonstrated high reliability:

- Description Agreement: 98.01%
- Visual Attributes Agreement: 98.56%
- Emotion Label Agreement: 91.47%

Data Statistics: The dataset includes significant representation of non-Western art, such as Ukiyo-e (Japanese) and Chinese Ink Wash painting. Analysis shows these "East Spirit" styles strongly correlate with "Low Arousal" and "Positive Valence" (Calmness), reflecting Eastern aesthetic ideals of harmony. Furthermore, EmoArt descriptions achieve a Type-Token Ratio (TTR) of 0.9358 and an Average Word Count of 16.22, significantly outperforming datasets like COCO and ArtEmis in lexical diversity.

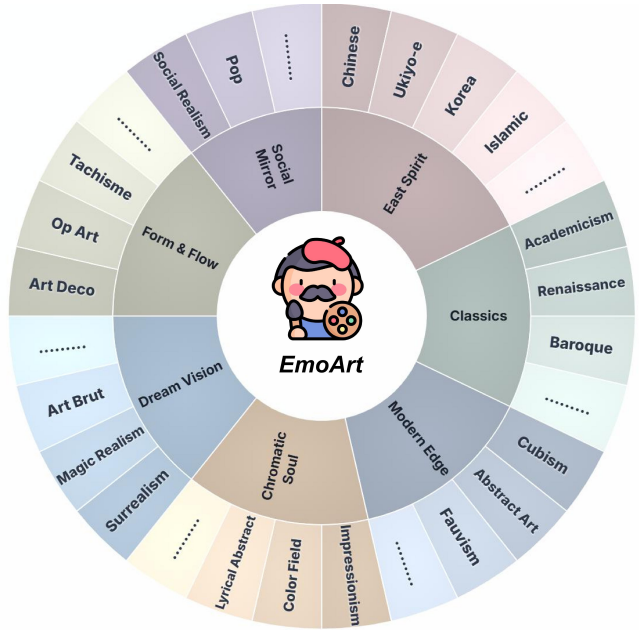

Figure 3: Representative art categories in the dataset: the inner ring shows the major categories, and the outer ring shows the specific subcategories.

3 Challenge Tasks

3.1 Track 1: Emotion-Aware Artistic Image Generation

Problem Statement: Given a multimodal prompt $P = \{T_{desc}, S_{style}, E_{target}\}$, where T_{desc} describes the semantic content, S_{style} specifies the artistic movement, and E_{target} defines the desired emotional state (via VA values or category), the goal is to generate an image I_{gen} that fulfills all three conditions.

Scientific Challenge: Participants must develop models that can disentangle style from emotion. For example, generating an "Expressionist" painting (Style) that is "Calm" (Emotion) is non-trivial, as Expressionism is typically associated with high arousal (Anger/Anxiety). Successful models must learn to subtly manipulate visual attributes (e.g., smoothing the brushwork while keeping the distorted forms) to achieve this hybrid state.

Submission Format: Participants will submit generated images for a hidden test set of prompts. Code/Docker containers are required for verification.

3.2 Track 2: Multidimensional Art Emotion Understanding

Problem Statement: This is a multi-task art-affect understanding challenge. Given an input artwork I, a model M is expected to produce a compact but information-rich report that covers both *what* is depicted and *how* the depicted content is emotionally conveyed through artistic techniques:

(1) **Emotion Prediction**: A probability distribution over the 12 emotion classes.
(2) **Binary VA Prediction**: Binary prediction for Valence (positive vs. negative) and Arousal (high vs. low).
(3) **Attribute Analysis**: Textual descriptions of the five visual attributes (Brushwork, Composition, Color, Line, Light).

Scientific Challenge: The key difficulty lies in connecting low-level visual patterns (e.g., texture, contrast, stroke direction) to high-level affective judgments that are subjective and context-dependent. In other words, the model should not only output an emotion label, but also explain it through interpretable attribute evidence, moving toward automated, controllable "Art Critics."

4 Baseline Models

We provide high-quality baselines to facilitate participation.

4.1 Generation Baseline: Fine-Tuned FLUX.1-dev

We utilized the FLUX.1-dev model, a state-of-the-art 12B parameter rectified flow transformer.

- **Fine-tuning**: We employed LoRA (Low-Rank Adaptation) for efficient fine-tuning [5].
- **Training Set**: A curated subset of 50 images per style (approx. 2,800 images).
- **Prompt Construction**: "Style: [S]. Arousal: [A]. Valence: [V]. Description: [T]."
- **Results**: In our preliminary benchmarks, this fine-tuned model achieved the highest scores in Brushstroke Quality (0.6388) and Color Quality (0.6974), significantly outperforming vanilla SDXL [12] and PixArt-sigma [2]. This demonstrates the necessity of fine-tuning on domain-specific emotional data.

4.2 Understanding Baseline

We provide a baseline using Qwen3-VL [13], a strong vision-language model that can be prompted to perform multi-head prediction and produce faithful natural-language rationales.

- **Input**: The artwork I (optionally with task instructions / few-shot examples).
- **Outputs**: (i) emotion classification, (ii) binary VA prediction, and (iii) attribute analysis as structured text that can be directly parsed into the five attribute slots.

5 Evaluation Protocol

The evaluation protocol is designed to rigorously measure **artistic fidelity, emotional alignment, and generalization ability** in a fair and leakage-resistant manner. Rather than relying solely on low-level pixel similarity, we prioritize metrics that capture semantic intent and affective consistency. The challenge consists of two complementary tracks: Track 1 evaluates emotion-aware artistic generation, and Track 2 evaluates emotion understanding. All official rankings are computed on a **hidden test set** evaluated exclusively on a secure server to preserve benchmark integrity.

5.1 Hidden Test Set and Evaluation Integrity

To prevent data leakage and ensure fair comparison, the official test set used for leaderboard ranking is a **hidden evaluation set that**

is not publicly released online. All test images are stored exclusively on the evaluation server and are inaccessible to participants.

Submissions are evaluated through a server-side pipeline in which participants upload executable inference code or prediction files, and results are computed internally against the hidden test data. This design discourages memorization-based approaches and prevents overfitting to test samples. The hidden test set is distributionally distinct from the public training data and contains diverse artistic styles and sources to assess true cross-domain generalization.

A final blind evaluation phase is conducted using a reserved subset of hidden samples that are never used for intermediate leaderboard updates. This protocol ensures the integrity, reproducibility, and scientific validity of the challenge results.

5.2 Track 1: Generation Metrics

Track 1 evaluates how well generated artworks preserve visual realism, stylistic intent, and requested artistic attributes. We employ a combination of perceptual and semantic metrics:

(1) **Fréchet Inception Distance (FID).** We report FID as a standard measure of distributional similarity between generated and reference artworks [4]. FID captures both image quality and diversity at a global level and serves as a common baseline for comparison with prior generative models.

(2) **Attribute Alignment Score (AAS).** We introduce the **Attribute Alignment Score (AAS)** to evaluate whether generated images faithfully realize the requested artistic attributes (e.g., style, brushwork, composition). A multimodal evaluator based on MiniCPM-V-2.6 [17], fine-tuned on EmoArt annotations, produces a textual attribute description $\hat{T}_{attr}$ for each generated image. AAS is defined as

$$AAS = \mathrm{CosSim}\big(\mathrm{CLIP}(\hat{T}_{attr}), \mathrm{CLIP}(T_{attr})\big), \tag{1}$$

where T_{attr} denotes the ground-truth attribute specification. This metric explicitly measures semantic alignment between intended and generated artistic techniques.

(3) **LPIPS Perceptual Similarity.** We report LPIPS [20] to assess perceptual realism and structural coherence at a local level. Together with FID, LPIPS provides a complementary view of visual fidelity and diversity.

The final Track 1 ranking is computed using a weighted aggregation of normalized scores to balance perceptual quality and semantic alignment.

5.3 Track 2: Emotion Understanding Metrics

Track 2 evaluates emotion recognition from artistic images along three core dimensions: categorical emotion, valence, and arousal. To ensure comparability and reproducibility, the official leaderboard is restricted to these standardized indicators.

- **Emotion Classification (12-way).** We report Top-1 Accuracy as the primary metric and Macro F1-score to account for class imbalance. This measures fine-grained recognition of discrete emotional categories expressed in artworks.
- **Valence Prediction (binary).** We report Accuracy and Macro F1-score for valence (positive vs. negative affect). This evaluates whether models capture the global emotional polarity conveyed by artistic cues.
- **Arousal Prediction (binary).** We report Accuracy and Macro F1-score for arousal (high vs. low activation). This measures sensitivity to intensity-related visual signals such as contrast, color saturation, and dynamic brushwork.

5.4 Human Expert Evaluation (Final Phase)

For top-ranked submissions, we conduct an additional qualitative evaluation performed by **human experts with formal training in art and visual design**. Evaluators are selected from individuals with academic or professional backgrounds in fine arts, art theory, or visual communication to ensure informed judgment of artistic quality and emotional expression.

Experts assess artworks using structured criteria including emotional expressiveness, artistic coherence, and technique realism (e.g., plausibility of brushwork and texture). Each sample is rated by multiple independent evaluators using standardized guidelines, and inter-rater agreement is monitored to ensure reliability.

This expert evaluation complements automatic metrics by capturing nuanced artistic and affective qualities that are difficult to quantify computationally, providing a holistic assessment of emotionally intelligent artistic generation and understanding.

Practical Notes: We will provide an evaluation script with a fixed output format and publish per-style breakdown tables (e.g., Impressionism vs. Abstract Expressionism) to highlight generalization across artistic movements. We also encourage (but do not require) participants to submit auxiliary analyses (e.g., calibration curves or qualitative case studies) in their technical reports to better explain strengths and failure modes.

6 Tentative Schedule

The challenge will follow the overall ACM Multimedia 2026 timeline and the Grand Challenge organization process. To keep the organization flexible and avoid unnecessary constraints, we provide the following *tentative* milestones as a reference; the exact dates and details may be adjusted based on conference announcements, platform readiness, and community feedback.

- **Early phase (launch & onboarding)**: We will open the challenge website, publish participation guidelines, and release an initial version of the training/validation resources together with baseline models.
- **Middle phase (development & evaluation)**: We will progressively release evaluation instructions and (when ready) the test protocol. During this period, we will also maintain a public FAQ and update baselines as needed to ensure reproducibility.
- **Late phase (submission & reporting)**: We will collect submissions, verify results, and announce the final leaderboard. A technical report/paper submission window will be provided according to the official proceedings requirements.
- **Workshop & dissemination**: Results will be presented and discussed in the associated workshop/session at ACM Multimedia 2026, including invited talks, participant presentations, and a summary of lessons learned.

Note: Any schedule updates will be communicated through the challenge website and official channels to ensure transparency for all participants.

7 Administrative Details

- **Website & Hosting**: The challenge will be hosted on a dedicated website (GitHub Pages) linked to a CodaLab competition page for automated leaderboard updates.
- **Data Access**:The dataset is hosted on Hugging Face (already accessible at https://huggingface.co/datasets/printblue/EmoArt-130k) under a CC BY-NC 4.0 license. Participants must sign an End User License Agreement (EULA) ensuring the data is used for academic research only.
- **Ethical Considerations**:
  - **Copyright**: All images are sourced from the public domain or open access museum policies.
  - **Bias**: We acknowledge that art history contains inherent biases (e.g., gender representation). We encourage participants to perform bias analysis on their models.
  - **Misuse**: The generation of deepfakes or offensive content is strictly prohibited. The organizing committee reserves the right to disqualify submissions violating these ethical guidelines.

8 Organizers

Hongxia Xie [Main Contact] is an Associate Professor at the College of Computer Science and Technology, Jilin University, China, and leads the Affective Vision Computing Lab. Her research focuses on computer vision, affective computing, and multimodal large models, with an emphasis on emotion understanding and generation in visual media. She has led multiple research projects in multimodal affective intelligence and actively promotes interdisciplinary collaboration between academia and industry. She is the corresponding author of the EmoArt dataset. She has published over 20 papers in leading international venues, including CVPR, ICCV, ECCV, ACM Multimedia, AAAI, and *IEEE Transactions on Affective Computing*. She also serves as the Grand Challenge Chair of ACM Multimedia 2025. (Email: hongxiaxie@jlu.edu.cn)

Cheng Zhang [Main Contact] is a research assistant working on multimedia understanding and generative models, with interests in vision–language learning and affective computing. His work focuses on fine-grained visual analysis and emotion-aware multimedia generation. He is the leading author of the EmoArt dataset. (Email: zhangcheng2122@mails.jlu.edu.cn)

Wen-Huang Cheng is a University Distinguished Chair Professor in the Department of Computer Science and Information Engineering at National Taiwan University and a Visiting Professor at the Korea Advanced Institute of Science and Technology (KAIST). His current research interests include multimedia, computer vision, and machine learning. He has actively participated in international events and played significant leadership roles in prestigious journals, conferences, and professional organizations. These roles include serving as Editor-in-Chief for IEEE CTSoc News on Consumer Technology, Senior Editor for IEEE Consumer Electronics Magazine (CEM), Associate Editor for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and IEEE Transactions on Multimedia (TMM), General Chair for ACM MMAsia (2023), IEEE ICME (2022), and ACM ICMR (2021), Technical Program Chair for ACM MM (2025), ACM ICMR (2022), IEEE ICME (2020), IEEE VCIP (2018), Chair for IEEE CASS Multimedia Systems and Applications (MSA) technical committee, and governing board member for IAPR. He has received numerous research and service awards, including the NVIDIA Academic Grant Program Award (2025), the 2024 Best Paper Award of IEEE Consumer Electronics Magazine, the Best Paper Award at the 2021 IEEE ICME and the Outstanding Associate Editor Award of IEEE TMM (2021 and 2020, twice). He is an IEEE Fellow, IET Fellow, and ACM Distinguished Member.

Jianlong Fu is a Principal Research Manager leading research and innovation in the Multimedia Computing Group at Microsoft Research Asia (MSRA), Beijing, China. He received his Ph.D. from the Institute of Automation, Chinese Academy of Sciences. His research focuses on multimedia content understanding and multimodal perceptual computing across images, videos, and embodied agents. He has published over 200 peer-reviewed papers and holds more than 20 U.S. patents, with a Google Scholar h-index of 63. Dr. Fu serves as Vice-Chair of the Automotive CE Applications Technical Committee under the IEEE Consumer Technology Society and is on the editorial boards of *IEEE Consumer Electronics Magazine* and *IEEE Transactions on Multimedia*. He also served as Guest Editor for *IEEE TPAMI* (2019–2021). He has chaired multiple committees at flagship conferences including ACM Multimedia (2021, 2026) and ACM ICMR (2021, 2023). His honors include the ACM SIGMM Rising Star Award 2022, the Best Paper Award at ACM Multimedia 2018, and more than ten international competition championships at CVPR/ICCV/ECCV. His research has been deployed in major Microsoft products including Windows, Office, Bing, Edge, and XiaoIce.

Sicheng Zhao is an Associate Professor at Tsinghua University. He was a postdoctoral research scientist at Columbia University (2020–2022), a postdoctoral research fellow at the University of California, Berkeley (2017–2020), and a postdoc at Tsinghua University (2016–2017). He received his Ph.D. from Harbin Institute of Technology in 2016 and was a visiting researcher at the National University of Singapore (2013–2014). His research focuses on affective computing, multimodal large models, and computer vision. He has published over 100 papers in top-tier venues with more than 14,000 Google Scholar citations. He is a Distinguished Member of CCF and a Senior Member of ACM, IEEE, and CSIG. He serves as Associate Editor for *IEEE TIP* and *IEEE TAFFC*, and has been Area Chair for NeurIPS, ICML, ICLR, ACM MM, CVPR, and ECCV. He has organized special sessions at ICASSP, ICMR, ICME, and ICIP, and workshops at ACM MM. He has been recognized as an AI 2000 Most Influential Scholar and a World's Top 2% Scientist.

Sanghoon Lee is an IEEE Fellow and Full Professor in the Department of Electrical and Electronic Engineering at Yonsei University, and an Adjunct Professor at the Yonsei University College of Medicine. His research interests include generative AI and deepfake detection, 3D multi-camera systems and reconstruction, human avatar modeling and synthesis, and multimodal signal processing

for artificial intelligence. He also served as an editor for the Journal of Communications and Networks from 2009 to 2015. He was an associate editor and guest editor of the IEEE Transactions on Image Processing from 2010 to 2014, and 2013, respectively. He was the General Chair of the 2013 IEEE IVMSP Workshop. He has served as the Chair of the IEEE P3333.1 Working Group since 2011. He served as an associate editor and a senior area editor for the IEEE Signal Processing Letters from 2014 to 2018, and since 2018. He was a member of the IEEE IVMSP/MMSP TC from 2014 to 2019 and from 2016 to 2021, respectively. He has served as an associate editor of the IEEE Transactions on Multimedia since 2022. He was the Image, Video, and Multimedia TC Chair of APSIPA from 2018 to 2019. He is a BoG member of APSIPA and is the Editor-in-Chief of APSIPA newsletters.

Xing Huang is the Executive Dean of the Insight Research Institute at Lianxin Digital, where she leads the development of human-centric AI models for large-scale psychological assessment. Her research focuses on quantifying transient psychological states and stable traits, delivering scalable affective computing solutions for national security, social stability, and public health. Her technologies have been commercially deployed in over 20 provinces in China and serve more than 400 institutional clients. She has contributed to multiple industrial standards in AI-driven psychological assessment and published in venues such as *Acta Psychologica*, *Scientia Sinica Vitae*, and ACM Multimedia.

Ling Lo is a postdoctoral researcher at National Yang Ming Chiao Tung University, Taiwan. She received her Ph.D. and B.S. in Electronics Engineering from National Yang Ming Chiao Tung University. Her research lies at the intersection of computer vision and machine learning, with a focus on affective computing, generative AI for multimedia, and trustworthy visual content creation. Her work has been published in top-tier venues, including ACM MM, CVPR, ICCV, AAAI, ACM TOMM, IEEE TMM, and IEEE TAFFC, and she received the ICME Best Paper Award (2021).

Hong-Han Shuai is a Professor at National Yang Ming Chiao Tung University (NYCU), where he is at the forefront of research in multimedia processing, deep learning, computer vision, and data mining. His research has been prominently featured at leading Artificial Intelligence/Data Mining conferences such as MM, NeurIPS, CVPR, ICCV, ECCV, ACL, EMNLP, NAACL, AAAI, KDD, WWW, ICDM, CIKM, and VLDB, and in top-tier journals including TKDE, TKDD, TMM, TNNLS, TACL, and JIOT. He has played significant roles in the academic community, serving as Associate Editor for both IEEE Transactions on Multimedia (TMM) and IEICE Transactions on Information and Systems, Guest Editor for ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), and as TPC Co-Chair for conferences like IEEE ICCE-TW, International Conference on Technologies and Applications of Artificial Intelligence (TAAI) and IPPR Conference on Computer Vision, Graphics, and Image Processing (CVGIP). He also contributes to the technical committee for IEEE Multimedia Systems and Applications (MSA). He has been honored with several research awards, notably the Best Paper Awards at the 2021 IEEE International Conference on Multimedia and Expo, 2025 World Congress on Medical and Health Informatics (MedInfo).

Chieh-Yun Chen is currently a Ph.D. student in Georgia Tech, advised by Prof. Humphrey Shi, where she focus on user-centric tasks within Vision-related fields, i.e., agent, generative models.

Ziyun Li is a Postdoctoral Researcher in Generative AI at KTH Royal Institute of Technology, Stockholm, Sweden. Her research interests include generative AI, flow matching, diffusion models, fairness, AI safety, and the societal impact of AI.

Jian-Yu Jiang-Lin is currently pursuing the Ph.D. degree with the Communications and Multimedia Laboratory, National Taiwan University, Taipei, Taiwan. He received the B.S. degree in electronics engineering and the M.S. degree in artificial intelligence from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2022 and 2024, respectively. His research interests include trustworthy Multimodal Large Language Models, with a particular focus on safety alignment, explainability, deepfake detection, and embodied intelligence.

Kang-Yang Huang received the M.S. degree in Computer Science from National Taiwan University (NTU), Taiwan, in 2025, and he earned the B.S. degree from the Undergraduate Honors Program in Electrical Engineering and Computer Science at National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. His research focuses on the intersection of computer vision and deep learning, encompassing generative AI and visual language models. He has received the 2024 Best Paper Award from IEEE Consumer Electronics Magazine.

Ling Zou is currently pursuing the Ph.D. degree in Computer Science at National Taiwan University (NTU), Taipei, Taiwan. She received the M.S. degree in computer science and information engineering from NTU in 2025, and the B.S. degree in computer science and information engineering from Ming Chuan University (MCU), Taiwan. Her research focuses on the intersection of computer vision and deep learning, with particular interests in the acceleration of multimodal large language models (MLLMs) and generative AI.

8.1 Team Composition, Division of Roles, and Geographic Diversity.

Our organizing team integrates complementary strengths from academia and industry while reflecting strong international and geographic diversity. The academic core consists of leading researchers in computer vision and multimedia — **Hongxia Xie** (China), **Wen-Huang Cheng, Hong-Han Shuai, Ling Lo** (Taiwan), **Sicheng Zhao** (China), **Sanghoon Lee** (South Korea), and **Ziyun Li** (Sweden) — who contribute expertise in affective computing, multimodal understanding, and generative AI. They are primarily responsible for defining the scientific scope, benchmark design, and academic program development.

The team is further strengthened by distinguished **industry** leaders, **Jianlong Fu** (Microsoft Research Asia, China) and **Xing Huang** (Lianxin Digital, China), who represent leading enterprise research and large-scale commercial deployment of AI technologies. They bring extensive experience in translating cutting-edge research into production systems, driving technology transfer, and shaping industrial standards. Their involvement explicitly anchors the workshop in real-world industrial scenarios, ensuring that the

program highlights enterprise-driven challenges, practical evaluation settings, and dedicated industry-focused sessions. In particular, **Lianxin Digital** will serve as a close industrial partner of the competition and provide sponsorship support, further strengthening the connection between academic research and real-world deployment. This strong industrial presence guarantees high practical relevance and fosters direct interaction between academic advances and practical applications.

This collaboration spans multiple countries and regions across Asia and Europe and bridges academic and industrial ecosystems. Such geographic and institutional diversity promotes inclusive participation, broad international outreach, and a balanced perspective that advances both fundamental research and impactful real-world innovation.

8.2 Commitment to Long-Term Website Maintenance

The organizing committee is fully committed to establishing, publishing, and continuously maintaining an official website for the AffectiveArt Grand Challenge. The website will provide comprehensive and up-to-date information, including challenge descriptions, task definitions, evaluation protocols, datasets (subject to license), baseline implementations, submission guidelines, leaderboards, and final results.

We commit to maintaining and updating the website, datasets, and related documentation for at least three years following the conclusion of ACM Multimedia 2026. The website will serve as a stable and publicly accessible platform to ensure long-term availability of challenge resources, reproducibility of results, and sustained community engagement.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11569–11579.

[2] Junsong Chen et al. 2024. PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. arXiv:2403.04692 [cs.CV]

[3] Elise S Dan-Glauser and Klaus R Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43 (2011), 468–477.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[6] Black Forest Labs. 2024. Flux.1 AI. https://flux1ai.com/. Accessed: 2025-05-28.

[7] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*. 83–92.

[8] Laurent Mertens, Elahe Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. 2024. Findingemo: An image dataset for emotion recognition in the wild. *Advances in Neural Information Processing Systems* 37 (2024), 4956–4996.

[9] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior research methods* 37 (2005), 626–630.

[10] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision*. 579–595.

[11] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 860–868.

[12] Dustin Podell et al. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[13] Qwen Team. 2025. Qwen3-VL Technical Report. Technical report / preprint.

[14] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. doi:10.1037/h0077714

[15] Stability AI. 2024. Stable Diffusion 3.5 Large Model Card. https://huggingface.co/stabilityai/stable-diffusion-3.5-large. Accessed: 2025-05-28.

[16] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20383–20394.

[17] Yuan Yao et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800 [cs.CV]

[18] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[19] Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. 2025. EmoArt: A Multidimensional Dataset for Emotion-Aware Artistic Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 12644–12650. doi:10.1145/3746027.3758201

[20] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.