THE ART OF AI: HOW A MULTIMODAL MODEL REVEALS THE SECRETS OF HUMAN CREATIVITY IN PAINTINGS

Zhehan Zhang[1], *Meihua Qian*[1], *Li Luo*[1], *Ripon Kumar Saha*[2]

[1]Clemson University, SC, [2]Arizona State University, AZ

ABSTRACT

Assessing artistic creativity has long been a challenge. Traditional tests are widely used but often require time-consuming manual scoring. Thus, researchers are exploring a new way, such as machine learning (ML), to automate the assessment. Recent research on visual artistic creativity assessment has demonstrated that ML methods are effective but constrained by their reliance on visual data alone. This study integrates textual descriptions alongside visual data for a more holistic assessment of paintings' creativity, which is more sophisticated to measure than simple sketches. The multimodal model was fine-tuned and leveraged both visual and textual inputs. It achieved approximately 95.3% accuracy in predicting the painting creativity scores, demonstrating a strong positive correlation (Pearson r = 0.96) with expert ratings. The study allows a text-image evaluation of paintings' creativity to better align with human interpretations.

Index Terms— visual artwork, creativity, automated creativity scoring, machine learning, multimodal model

1. INTRODUCTION

Creativity is increasingly recognized as a critical skill in the 21st century, essential for driving innovation, addressing complex problems, and fostering social skills [1]. It is considered a cornerstone of education and a key ability in preparing individuals for future workforce demands [2]. Creativity also occupies a prominent position in frameworks such as Bloom's taxonomy and the Program for International Student Assessment (PISA), highlighting its value as a higher-order cognitive skill [3, 4]. Despite its importance, assessing creativity, particularly artistic creativity, remains a major challenge due to its subjective and multidimensional nature [5].

1.1. Definition of Creativity

Creativity involves producing something that is both original and effective, and originality refers to the extent to which a solution or idea is novel, surprising, or unique [6]. In terms of creativity in art, it emerges from a complex interplay of the individual characteristics and process of the creator, the artifact or product they create, and the surrounding social-cultural context [7]. As famous psychologist Mihaly Csikszentmihalyi stated, art creativity is not only an internal spark; it is also heavily influenced by domain conventions and audience reception, such as paintings [8].

1.2. How Creativity Emerges

According to Guilford [6], the creative process commences with problem recognition and definition, a stage characterized by convergent thinking aimed at identifying a specific problem that needs a solution.

Next, idea generation employs divergent thinking—the capacity to develop numerous varied and novel solutions for an open-ended problem. A popular divergent thinking technique is brainstorming, which aims to foster a multitude of ideas without initial judgment.

After generating a wide range of relevant and original ideas, the creative problem solving process moves to evaluating and selecting the most creative ones. This stage, unlike idea generation, depends on convergent thinking skills.

Finally, solution validation evaluates the chosen solution to confirm that the solution successfully addresses the specified problem. This phase uses both convergent and divergent thinking, because it involves applying the solution and possibly refining it based on feedback [7].

Figure 1 shows the key steps of how creativity emerges based on Guilford's theory.

1.3. Creativity Assessment

Historically, creativity assessment has relied on human judgment. One well-known example is the Torrance Tests of Creative Thinking (TTCT) [9, 10], which ask participants to list unusual uses for an object or complete drawings based on abstract figural prompts, and responses are typically scored manually in terms of fluency, flexibility, originality, and elaboration. Similarly, the widely used Consensual Assessment Technique (CAT) involves experts evaluating creative products (e.g., artworks, stories) based on either established or context-specific criteria [11]. These methods, while considered gold standards in the field of creativity, have clear limitations: they are labor-intensive (requiring trained raters), time-

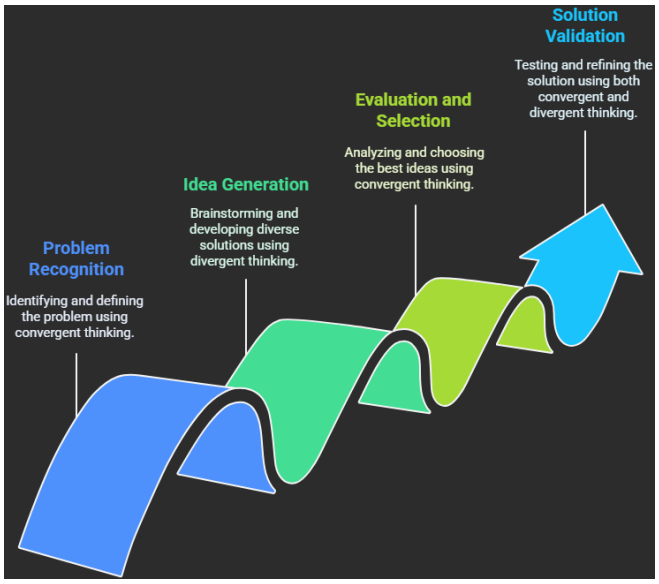

Fig. 1. The key steps of creativity emergence [6]

consuming to score, and often only feasible for small-scale evaluations [12]. Moreover, traditional tests like TTCT focus heavily on divergent thinking – for instance, how many novel ideas one can generate – which captures only part of the creative spectrum and divergent thinking (DT) scores may undervalue other aspects of creativity such as artistic style, emotional impact, or technical skill [13].

2. AI OR ML FOR CREATIVITY ASSESSMENT

Given the challenges associated with conventional creativity assessment methods, there is growing interest in leveraging advanced technologies, such as AI, ML and deep learning, to automate creativity assessment and develop more precise, objective, and efficient assessment tools. In the domain of visual arts, researchers have begun exploring whether computers can "judge" creativity in drawings or paintings in a manner consistent with human experts [14]. For example, Cropley and Marrone [12] developed an automatic scoring system using a convolutional neural network (CNN) model to classify drawings from high to low creativity. Patterson et al. [15] also used a CNN model, but incorporated a regression layer to generate continuous creativity scores for a larger sample dataset. A recent study by Acar et al. [5] suggests that a multimodal model-CLIP (Contrastive Language-Image Pre-training) reliable for scoring the creativity of drawings.

However, prior research on assessing creativity in visual art has focused on drawings. This study aims to assess the creativity of paintings through the development of a multimodal ML model. Figure 2 shows the difference between drawings in typical DT tests and a painting we study.

Fig. 2. Comparison of drawings in a typical DT figural creativity test and a random painting in this study [14].

2.1. What's new in our study?

This study specifically integrates CLIP's multimodal capabilities by incorporating textual descriptions of paintings alongside visual data. This approach enables a richer and more holistic assessment of human creativity in paintings. Within the creativity assessment area, this is one of the first efforts to evaluate artistic creativity—ranging from children's artwork to famous masterpieces—using a vision-language model. We extend previous image-only approaches by providing the model with contextual information through brief textual descriptions of each painting.

We also emphasize the cross-domain relevance of our method: it demonstrates how combining modalities can enhance the understanding of visual content more broadly. Similar multimodal strategies could improve tasks such as image captioning, visual question answering, or aesthetic quality assessment, where purely visual methods may overlook contextual nuances that text can provide. This work highlights the potential for AI systems that understand not only the "pixels" but also the "story" behind human creativity.

3. DATASET AND RUBRICS

The dataset in this study comprised 1,000 paintings curated from three distinct sources, the proportion of paintings in each category roughly matches expert-rated score ranges, respectively (0-75, 75-95, 95-100), with a wide range of creative content and high quality (minimum 600×600 pixels) : Children's Paintings (750 images): These were acquired through a licensed collection from iStock in high quality, using keywords like "kids painting" to retrieve amateur child art. The subjects in this subset tend to be simple and familiar. Professional Artists' Paintings (200 images) – Sourced from Wikimedia Commons, these paintings were created by various (often unknown or less famous) artists. We used a similar key-words search selection method, choosing the first 200 images meeting the resolution criteria. These works generally exhibit greater technical proficiency, detail, and complexity of composition. The subjects vary widely (portraits, landscapes, abstract concepts), providing richer visual features. Famous Masterpieces (50 images) – we incorpo-

rated 50 iconic paintings (e.g., Leonardo da Vinci's Mona Lisa, Van Gogh's Starry Night). These were sourced from public domain images. They represent high-creativity examples against which the model's sensitivity to creativity can be tested. Each painting is accompanied by a textual description. For children's and professional paintings, descriptions were provided by either the original source or generated by the researchers to summarize the content or artist intention. These texts range from simple content descriptions ("A child's drawing of a family in a house, with smiling stick figures") to more elaborate explanations of meaning or technique (especially for the professional paintings). For those famous works, brief art-historical notes or the artist's own commentary (when available) were used.

3.1. Expert Ratings and Rubric

Two experts in visual arts and creativity research independently rated all paintings using a standardized rubric, assigning 0–20 points in each of five dimensions:
Originality: Novelty, imagination, or unexpectedness; high scores for innovative concepts or stylistic originality.
Color: Use of color, harmony, contrast, and emotional impact; high scores for skillful or bold choices.
Texture: Perceived texture and technique (e.g., brushstrokes); high scores for complex or skillful texture.
Composition: Arrangement and design; high scores for balanced or intentionally unconventional layouts.
Content: Thematic depth or message; high scores for rich meaning, storytelling, or emotional impact.

Scores were summed (total 0–100), treating all dimensions equally, following the Consensual Assessment Technique [11] and Lu et al. [16]. Final creativity scores were the average of both experts' ratings. Inter-rater reliability was high (Intraclass Correlation Coefficient $= 0.99$), indicating strong consistency.

4. MULTIMODAL MODEL RATIONALE

We fine-tuned a Transformer-based multimodal regression model to predict the creativity scores of human paintings, leveraging both visual and textual inputs. Our model builds upon CLIP model by OpenAI [17], which learns a joint embedding space for image-text pairs through contrastive learning. Meanwhile, we augmented CLIP's textual embeddings with MiniLM sentence model [18] to enrich the textual representations with nuanced semantic features.

4.1. CLIP Model Overview

CLIP consists of two separate Transformer-based encoders: A Vision Transformer f_v that maps an image I_v to a dense embedding $v_i \in \mathbb{R}^d$. A Text Transformer f_t that maps a corresponding caption T_i to a text embedding $t_i \in \mathbb{R}^d$. The model is trained on a large-scale dataset of image-text pairs to maximize the similarity of matching pairs while minimizing the similarity of mismatched pairs in a shared embedding space. This is achieved using a contrastive loss, which for a batch of N image-text pairs is defined as Equation 1:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(v_i, t_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(v_i, t_j)/\tau)} \quad (1)$$

where $\cos(,)$ is cosine similarity, τ is a temperature hyperparameter. This bidirectional loss ensures alignment between images and their corresponding texts while maintaining separation from unrelated samples.

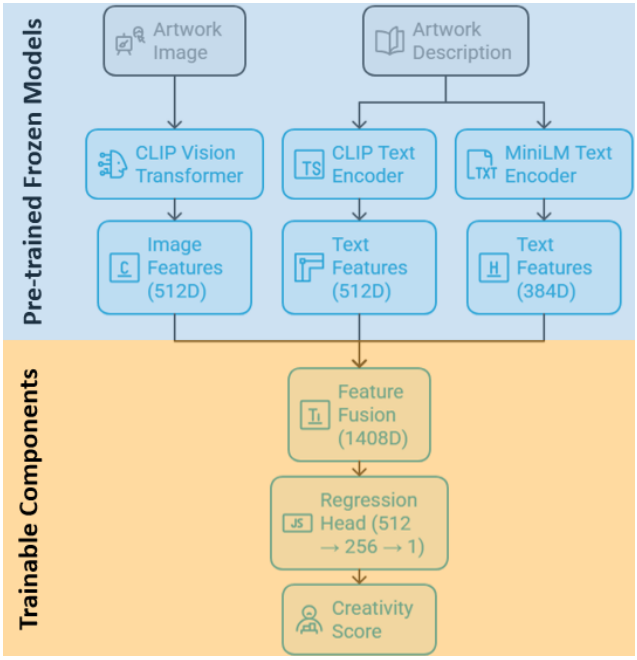

Fig. 3. Multimodal model: CLIP image and text (MiniLM) embeddings are fused for creativity prediction.

4.2. Data Preprocessing

All images were resized to 336×336 pixels for CLIP, with random augmentations (cropping, flips, rotations) applied during training to boost variability and reduce overfitting. Text descriptions were cleaned, tokenized with CLIP's tokenizer, and encoded into two vectors: $t_{CLIP} \in \mathbb{R}^{512}$ (CLIP text encoder output) and $t_{MiniLM} \in \mathbb{R}^{384}$ (MiniLM sentence embedding). These were concatenated to form the final text feature $t_i \in \mathbb{R}^{896}$ for each painting.

4.3. Multimodal Feature Fusion

As shown in Figure 3, each image I_i is encoded to a 512-d vector v_i by the CLIP vision encoder. The text feature

t_i (CLIP+MiniLM, 896-d) is concatenated with v_i to form $f_i = [v_i; t_i] \in \mathbb{R}^{1408}$. This fused vector passes through a small feed-forward network (two linear layers) to predict the creativity score.

4.4. Training and Validation

The overall model was trained end-to-end using Mean Squared Error (MSE) loss $L_{CLIP} = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{y}_i]^2$ where y_i is the ground-truth creativity score and $\hat{y}_i$ the model's prediction. Optimization was performed using the AdamW optimizer (learning rate = 1e-5, batch size = 16). The dataset was split into 80% for training and 20% for testing. To prevent overfitting given the relatively small dataset, we employed a 20% dropout on the final embedding layer and early stopping if validation loss (via a 10-fold cross-validation on training set, given no separate validation split) did not improve for 5 consecutive epochs. We verified that the model did not overfit by observing training vs. cross-validation error convergence (training MAE 4.5, cross-val MAE 5.0 at best epoch).

5. RESULTS

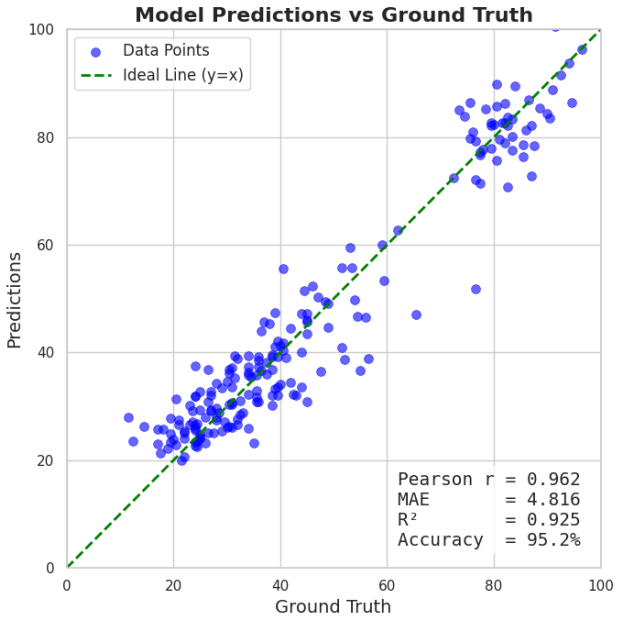

Fig. 4. The multimodal model predicted creativity scores vs. human expert scores for 200 test paintings

After training, the predictions closely matched the expert-provided creativity scores. Figure 4 shows a scatter plot of the model's predicted scores vs. the actual human scores for all 200 test paintings, with a green dot line indicating the ideal $y = x$ (perfect agreement). The points cluster tightly around the diagonal, reflecting a exceedingly strong correlation.

Quantitatively, the model achieved an average Pearson correlation of $r = 0.962$ with the ground-truth scores ($p < 0.001$). This indicates a strong linear relationship – (during the training Best Pearson Correlation=0.965). The highest coefficient of determination (R^2) is about 0.93, meaning the model's scores can explain ∼93% of the variance in human scores. It also reported the average Mean Absolute Error (MAE) on test predictions: 4.8 points on the 0–100 scale, which means on average the model was off by less than 5 points in either direction. Given the inherent subjectivity (even human judges often differ by a few points), an average error of ∼4.8 is quite low. If we consider a tolerance window – within ±5 points is accurate – the model's score falls in that window about 95% of the time, effectively 95% accuracy within 5 points. Comparing to the baseline, Table 1 shows the multimodal model outperformed the visual-only CNN model.

Table 1. Performance comparison of multimodal model against the CNN model when training on 1000 paintings.

Method	MAE (↓)	P-r (↑)	Accuracy (↑)
CNN	5.65	0.946	94.3%
CLIP Image+Text	5.05	0.959	94.9%
Multimodal	4.74	0.965	95.3%

To further analyze the contribution of each modality, we conducted an ablation study. Table 2 summarizes the results for different combinations of image, text, and MiniLM features.

Table 2. Ablation study results for different modality combinations. **Bold** indicates best results. ✓ denotes inclusion of a modality.

Input Modality			**Performance**		
Image	**Text**	**MiniLM**	**MAE (↓)**	**Pearson (↑)**	**Acc. (↑)**
✓			5.47	0.952	94.5%
	✓		8.13	0.899	91.9%
		✓	8.39	0.889	91.6%
✓	✓		5.05	0.959	94.9%
✓		✓	5.20	0.959	94.8%
	✓	✓	7.82	0.901	92.2%
✓	✓	✓	**4.64**	**0.965**	**95.3%**

6. CONCLUSION

Our multimodal model, fusing visual and textual inputs, accurately assesses painting creativity (r=0.965), proving that this synergy is vital for achieving a human-aligned understanding of art. This work provides a scalable tool for art analysis and moves AI toward a more nuanced appreciation of human creativity.

7. REFERENCES

[1] Branden Thornhill-Miller, Anaëlle Camarda, Maxence Mercier, Jean-Marie Burkhardt, Tiffany Morisseau, Samira Bourgeois-Bougrine, Florent Vinchon, Stephanie El Hayek, Myriam Augereau-Landais, Florence Mourey, et al., "Creativity, critical thinking, communication, and collaboration: assessment, certification, and promotion of 21st century skills for the future of work and education," *Journal of Intelligence*, vol. 11, no. 3, pp. 54, 2023.

[2] Seyedahmad Rahimi, Jason Brent Smith, Erin JK Truesdell, Ashvala Vinay, Kristy Elizabeth Boyer, Brian Magerko, Jason Freeman, and Tom Mcklin, "An automated, unobtrusive, formative assessment of creativity in a computer science and music remixing learning environment.," *Psychology of Aesthetics, Creativity, and the Arts*, 2024.

[3] Lorin W Anderson and David R Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*, Addison Wesley Longman, Inc., 2001.

[4] OECD., "Pisa 2022 results (volume i): The state of learning and equity in education," 2023.

[5] Selcuk Acar, Peter Organisciak, and Denis Dumas, "Automated scoring of figural tests of creativity with computer vision," *The Journal of Creative Behavior*, vol. 59, no. 1, pp. e677, 2025.

[6] J Guilford, "Creativity. american psychology. 5 (9), 444–454," 1950.

[7] David H Cropley, RL Marrone, K Medeiros, and K van Broekhoven, "Creative products and artificial intelligence," in *Creations: The Nature of Creative Products in the 21st Century*, pp. 33–59. Springer, 2025.

[8] Mihaly Csikszentmihalyi, *The systems model of creativity*, Springer, 2014.

[9] E Paul Torrance, "Torrance tests of creative thinking," *Educational and psychological measurement*, 1966.

[10] EP Torrance and ZL Rockenstein, "Styles of thinking and creativity," *Gifted International*, vol. 4, no. 1, pp. 37–49, 1986.

[11] Teresa M Amabile, "Social psychology of creativity: A consensual assessment technique.," *Journal of personality and social psychology*, vol. 43, no. 5, pp. 997, 1982.

[12] David H Cropley and Rebecca L Marrone, "Automated scoring of figural creativity using a convolutional neural network.," *Psychology of Aesthetics, Creativity, and the Arts*, 2022.

[13] Jonathan A Plucker, Matthew C Makel, and Meihua Qian, "Assessment of creativity," *The Cambridge handbook of creativity*, pp. 48–73, 2010.

[14] Zhehan Zhang, Meihua Qian, Li Luo, Ripon Saha, Qianyi Gao, and Xinxin Song, "Using a cnn model to assess visual artwork's creativity," *arXiv preprint arXiv:2408.01481*, 2024.

[15] John D Patterson, Baptiste Barbot, James Lloyd-Cox, and Roger E Beaty, "Audra: An automated drawing assessment platform for evaluating creativity," *Behavior research methods*, vol. 56, no. 4, pp. 3619–3636, 2024.

[16] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[18] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.