ArtMentor: AI-Assisted Evaluation of Artworks to Explore Multimodal Large Language Models Capabilities

Chanjin Zheng
chjzheng@dep.ecnu.edu.cn
Shanghai Institute of Artificial
Intelligence for Education, East China
Normal University
Shanghai, China
Faculty of Education, East China
Normal University
Shanghai, China

Zengyi Yu
Faculty of Education, East China
Normal University
Shanghai, China
202105720431@zjut.edu.cn
College of Education, Zhejiang
University of Technology
Hangzhou, Zhejiang, China

Yilin Jiang
zjut_jiangyilin@163.com
College of Education, Zhejiang
University of Technology
Hangzhou, Zhejiang, China

Mingzi Zhang
Faculty of Education, East China
Normal University
Shanghai, China
windyday@zjnu.edu.cn
College of Education, Zhejiang
Normal University
Jinhua, Zhejiang, China

Xunuo Lu
13968860822@163.com
School of Economy, Zhejiang
University of Technology
Hangzhou, Zhejiang, China

Jing Jin
School of Education, Zhejiang Normal
University
Jinhua, Zhejiang, China
383230730@qq.com
Tianchang Guanchao Primary School
Hangzhou, Zhejiang, China

Liteng Gao
2335060610@st.usst.edu.cn
School of Artificial Intelligence
Science and Technology, University of
Shanghai for Science and Technology
Shanghai, China

Abstract

Can Multimodal Large Language Models (MLLMs), with capabilities in perception, recognition, understanding, and reasoning, act as independent assistants in art evaluation dialogues? Current MLLM evaluation methods, reliant on subjective human scoring or costly interviews, lack comprehensive scenario coverage. This paper proposes a process-oriented Human-Computer Interaction (HCI) space design for more accurate MLLM assessment and development. This approach aids teachers in efficient art evaluation and records interactions for MLLM capability assessment. We introduce ArtMentor, a comprehensive space integrating a dataset and three systems for optimized MLLM evaluation. It includes 380 sessions from five art teachers across nine critical dimensions. The modular system features *entity recognition*, *review generation*, and *suggestion generation* agents, enabling *iterative upgrades*. Machine learning and natural language processing ensure reliable evaluations. Results confirm GPT-4o's effectiveness in assisting teachers in art evaluation dialogues. Our contributions are available at https://artmentor.github.io/.

CCS Concepts

- **Human-centered computing** → **Laboratory experiments**; *Human computer interaction (HCI)*; *HCI design and evaluation methods*.

Keywords

AI-Assisted Artwork Evaluation, GPT-4o, Multimodal Large Language Models, Human-Computer Interaction Dataset Design, Entity Recognition, Multi-Agent for Iterative Upgrades.

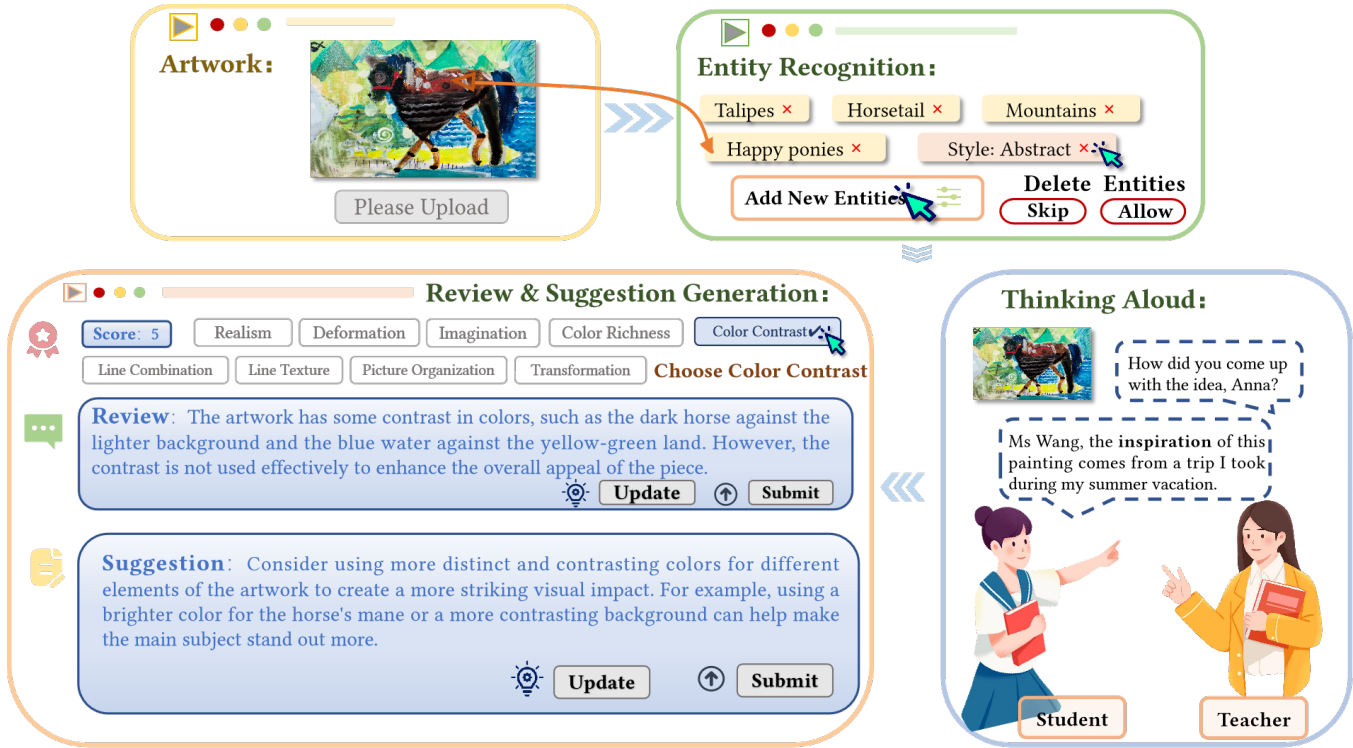*These authors contributed equally: Chanjin Zheng, Zengyi Yu, Yilin Jiang.
†Corresponding author: Chanjin Zheng (Email at chjzheng@dep.ecnu.edu.cn).

Figure 1: A multi-agent data collection system from ARTMENTOR specifically designed to assess the GPT-4o's assistance capabilities in art evaluation. It captures interactions across 380 evaluation sessions involving five art teachers and three agents of GPT-4o.

1 INTRODUCTION

Multimodal large language models (MLLMs), by seamlessly integrating various data types such as text and images, possess capabilities in multimodal perception, recognition, understanding, and reasoning [24]. Recent MLLMs, including GPT-4o [1], Gemini [51, 59], and Claude 3 [8], excel in tasks like image recognition, visual question answering, cross-modal retrieval, and video understanding [62, 65]. These strengths in visual analysis and language generation offer significant potential for advancing art education assessment [54, 69]. In the realm of art education, particularly artwork evaluation, the integration of the assessment process with broader dialogue-based education is pivotal [25, 30]. This approach enhances students' understanding and mitigates concerns of dehumanization. At the intersection of arts, education, and AI, a key question arises: *Can MLLMs function as independent entities in collaborative evaluation processes*, effectively supporting teachers and enriching educational experiences? In artwork evaluation, MLLMs can identify visual elements ("*entities*") such as trees, faces, and art styles [14]. Based on entity recognition results, MLLMs can generate artwork reviews and suggestions, functioning like an "*ArtMentor*". This raises a crucial question: **How effectively can MLLMs assist elementary art teachers with *entity recognition, review generation,* and *suggestion generation*?**

To assist elementary art teachers effectively, it is crucial to understand MLLMs' capabilities and limitations within specific HCI

contexts [11, 23]. This includes both their general performance and their behavior in educational settings [50]. MLLMs can identify entities in artworks, link them to historical art movements, and analyze color schemes or compositions [6, 29]. However, their effectiveness relies heavily on the HCI methods used [45]. In specialized HCI art environments like elementary art classrooms, the depth and accuracy with which MLLMs comprehend artistic elements are vital.

To better tailor MLLMs for educational purposes, we have streamlined the assistance process into three distinct phases. Each session starts with an uploaded artwork, followed by the MLLM's automatic recognition of entities within the artwork. Incorporating "Thinking Aloud" [21] and "Protocol Analysis" [22], our system also encourages students to create audio recordings that elucidate their creative ideas and critically analyze their own work for art teachers. Art teachers refine the entities, including art style, until accurate. After entity recognition, MLLMs review the artworks and assign scores. HCI methods verify the appropriateness of these scores, after which MLLMs provide suggestions for modifying the artwork. Each stage of this process is managed by a designated agent—namely, the *entity recognition agent, review generation agent*, and *suggestion generation agent*. This modular approach supports *iterative updates* and targeted enhancements, ensuring adaptability and effectiveness in educational applications. This multi-agent framework raises key questions: Can the *entity recognition agent* accurately identify the themes in student artwork? How effective are the *review generation*

agent and *suggestion generation agent* in enhancing the artistic process? Systematically addressing these questions will offer insights for refining our design. Without clear answers, assumptions about MLLMs' effectiveness in assisting teachers remain speculative.

To address these challenges, our study draws inspiration from Mina Lee's works [34, 35], particularly her focus on evaluating the writing capabilities of LLMs (GPT-3) through **HCI datasets** at CHI 2022, and her subsequent proposal of a **design space** for systematically exploring intelligent interactive writing assistants at CHI 2024. Building on this foundation, we adopt an MLLM-driven approach that collects process-oriented data during interactions between art teachers and MLLMs, expanding the scope to the multimodal domain and breaking down artistic evaluation capabilities into multiple sub-abilities.

This dynamic and ongoing data type authentically reflects the MLLM's capabilities by capturing the evolving interaction, thus avoiding the biases of result-oriented evaluations [28]. In HCI research, such process-oriented data is challenging to fabricate, offering a more reliable assessment of the model's performance in educational contexts. However, due to factors like context variability and decoding parameters [37, 68], testing these capabilities poses significant challenges in HCI settings. For example, after multiple interactions between an elementary school art teacher and MLLMs, repeatedly modifying and refining reviews and suggestions on student artworks, how can we accurately identify and describe the model's specific contribution to the final reviews and suggestions? How can we quantify the extent to which the MLLMs meet the art teacher's assistance needs?

The *design and analysis of HCI spaces*, encompassing datasets, multi-agent data collection systems and so on, are essential for addressing the evaluation challenges faced by MLLMs in educational assistance contexts. To gather raw data, we utilized GPT-4o, a representative MLLM, and developed the ARTMENTOR, as shown in Figure. 1. Specifically, the ARTMENTOR space allows for an in-depth analysis of GPT-4o's ability to assist teachers across nine key dimensions: realism, deformation, imagination, color richness, color contrast, line combination, line texture, picture organization, and transformation. Consequently, we organize both the review generation and suggestion generation agents into nine distinct *dimensions*

We demonstrate that the metrics designed to integrate HCI datasets with machine learning and natural language processing effectively quantify the assistance capabilities of MLLMs. Specifcally, we adapt standard machine learning metrics such as accuracy, precision, recall, and F1-score to evaluate entity recognition capabilities, detailed in section 4.2.1. Additionally, we introduce score acceptance Models outlined in section 4.2.3. Drawing from natural language processing, we assess the generation of reviews and suggestions using two criteria: text modification length and text similarity, as discussed in section 4.2.4. Lastly, we evaluate art style sensitivity to gauge the acceptance of art styles, which is elaborated in section 4.2.2.

This paper makes three contributions: **(1)** We introduce a multi-agent space named ARTMENTOR, which effectively collects process-oriented HCI datasets to mitigate the fabrication issues often found in result-oriented data. The dataset and code for this space are freely accessible at https://artmentor.github.io/; **(2)** We develop comprehensive evaluation metrics by integrating insights from machine learning, natural language processing, and HCI to holistically assess the assistance capabilities of MLLMs; **(3)** Through extensive data collection and analysis, We identified underperforming dimensions in the multi-agent system and proposed iterative enhancements to improve overall performance. *These contributions pave the way for a more refined, process-oriented, and versatile approach to the evaluation of MLLMs' capabilities.*

2 RELATED WORK

2.1 Capability Evaluation of MLLMs

2.1.1 Types of Capability. The exploration of the capabilities and limitations of MLLMs is crucial for designing effective multimodal interactions. Directly aligning with how MLLMs assist in *entity recognition*, *review generation* (including *score*), and *suggestion generation* for elementary art teachers, this foundational exploration is divided into four key sections [24]:

- **Multimodal Perception:** Examines MLLMs' understanding of spatial and relational dynamics within data from different modalities. This includes: (1) Object localization, which involves determining the position and orientation of objects within scenes, crucial for spatial awareness [13, 63]; (2) Object relation, identifying spatial and contextual relationships between objects [4, 41]; (3) Object interaction, recognizing interactions that involve actions, movements, or functional relationships within a visual context [12, 70]. *This multifaceted approach corresponds to our system's capability to facilitate holistic assessment (art style recognition), capturing the intricate interplay of various artistic elements.*

- **Multimodal Recognition:** Focuses on the identification and classification of entities, actions, and attributes across different modalities, which includes: (1) Concept recognition, assessing models' ability to categorize objects, actions, and scenes from varied sensory inputs [36, 41, 63]; (2) Attribute recognition, evaluating the detection of styles, emotions, and quantities across different modalities [3, 41]; (3) Action recognition, interpreting actions within various contexts [18, 41]; (4) Text recognition, determining the ability to transcribe text from images, vital for processes like automated documentation [1, 41]. *This aligns with our entity recognition process in the ARTMENTOR space.*

- **Multi-modal Understanding:** This section evaluates MLLMs on their capability to process and make sense of data from multiple sensory inputs, extending beyond textual information, to provide a comprehensive understanding of multimodal data integration [24]. *This principle aligns with our system's capability to generate in-depth reviews and nuanced scores for individual dimensions of art evaluation, demonstrating a profound understanding of specific artistic aspects.*

- **Multimodal Reasoning:** Investigates how MLLMs infer logical conclusions from multimodal data. This section covers: (1) Commonsense reasoning, which evaluates models' ability to apply knowledge to interpret interactions and relationships within images [39, 64]; (2) Relation reasoning, testing understanding of social, physical, or natural relations among various elements [42, 51]; (3) Logic reasoning, assessing the application of logical principles in analyzing and interpreting multimodal information [5, 41].

This is closely related to how we generate suggestions in the ART-MENTOR space based on the assessed dimensions.

2.1.2 Methods to Evaluate. The evaluation of MLLMs encompasses several methodologies that ensure a comprehensive evaluation of their capabilities. These methods are divided into three primary categories [24]:

- **Human Evaluation:** Human evaluators play a crucial role in assessing the capabilities of MLLMs, especially in tasks that demand high comprehension levels and are challenging to quantify using standard metrics. The evaluation focuses on multiple dimensions including: (1) Relevance, assessing whether the responses align with the intended instructions [40]; (2) Coherence, determining if the responses are logically structured and consistent; (3) Fluency, evaluating the naturalness and grammatical correctness of the generated outputs.
- **GPT-4 Evaluation:** To complement human evaluation and address its resource-intensive nature, the instruction-following capabilities of GPT-4 are used to efficiently evaluate model-generated outputs. GPT-4 assesses the MLLMs on dimensions such as helpfulness, relevance, accuracy, and detail, scoring them on a scale from 1 to 10, where higher scores indicate better performance. This approach not only provides scores but also detailed explanations for the evaluations, offering insights into the model's strengths and areas for improvement [1, 40].
- **Metric Evaluation:** While qualitative insights from human and GPT-4 evaluations are valuable, traditional metrics are essential for quantitatively assessing MLLM performance. These metrics provide standardized and objective measurements across various tasks: (1) For recognition capabilities, metrics like Accuracy and Average Precision are utilized [33, 36, 38]; (2) For perception capabilities, measures such as mIoU, mAP, and Dice are adopted [17]; (3) For evaluating text or image generation capabilities, metrics such as BLEU, ROUGE, and METEOR are widely employed [15, 31], providing clear indicators of a model's performance in various applications.

While human evaluation offers insightful perspectives, it is inherently subjective and costly, with GPT-4 assessments potentially varying due to fluctuations in prompts and parameters. Furthermore, Mina Lee and colleagues have deliberated on two methodologies for investigating the generative capacities of large language models (LLMs): contextual inquiry and interaction logging analysis. Contextual inquiry, through interviews, provides profound insights albeit with limited generalizability [35]; interaction logging analysis, though broad in scope, lacks depth. Previous research has largely been confined to specific tasks and settings. We aim to synthesize these approaches, customizing them for multi-modal tasks using MLLMs, thereby addressing the limitations of interviews and validating the model across a wider spectrum, offering a more comprehensive evaluation and insights for future research. Consequently, *we integrate traditional machine learning metrics with our innovative natural language processing techniques to deliver a nuanced, robust, and reliable assessment of MLLMs.*

2.2 Process-oriented HCI Datasets in Education

2.2.1 The Growing Focus on Educational Process Mining in HCI. In the educational domain, the process of learning is often considered more critical and analytically valuable than the final outcomes [61]. Capturing this process, however, poses significant challenges due to the complexity of documenting and analyzing process-oriented data, which tends to be sparse and less frequently analyzed. Common applications of process mining techniques have been demonstrated in online assessment data to analyze the processes involved in answering questions or requesting student feedback [48]. Additionally, frameworks integrating educational process data mining have been introduced to facilitate the handling of interactive process data and assist educators in analyzing educational processes based on formal modeling [60]. Despite these advancements, the collection and analysis of such data remain labor-intensive and inherently complex. The complexity of process-oriented data in educational settings correlates strongly with concepts such as simplicity, ease of use, uncertainty, and the context of application, making it a focal point for HCI designers [61]. These challenges underscore the need for innovative approaches in HCI to enhance the usability and effectiveness of process mining tools in educational environments. *Consequently, the significance and ongoing challenges of Educational Process Mining in HCI have garnered increasing attention, highlighting the urgency for developing more efficient and accessible tools.*

2.2.2 Emerging Trends in Process-Oriented Artwork Evaluation. In the field of artwork evaluation, process-oriented approaches are beginning to take shape. Currently, there are very few widely validated methods for automated process-oriented visual arts assessment. In fact, result-oriented methods are also scarce; one of the few examples is the Torrance Tests of Creative Thinking—Drawing Task, which scores creativity using artificial neural networks [16]. One emerging process-oriented method involves providing an initial artwork that allows students to further develop the piece by adding patterns. The evaluation is then conducted based on both the initial and final artworks [47]. This method attempts to document the creative process of students' artwork creation as much as possible, but recording the entire creative process remains a significant challenge. This raises an important question: *Can the process of artwork evaluation under AI assistance be effectively documented?* Through the design of ARTMENTOR, we aim to facilitate interaction between educators and the system, deepening their understanding of students' artworks, thereby advancing the development of HCI in the domain of art education.

2.3 Spaces for Iterative Multi-Agent Upgrades

2.3.1 Fractionalization and Dominance. The *CoAuthor* study has significantly contributed to the HCI community by highlighting the generative capabilities of large language models (LLMs) in creative and argumentative writing contexts [35]. While *CoAuthor* effectively advocates for the curation and analysis of large interaction datasets to make these capabilities more transparent and accessible, it does not segment the creative process into distinct phases that could provide deeper, context-specific insights. Building on these viewpoints, we propose that applying a similar phased approach to the art evaluation process—identifying distinct stages like conception, development, and presentation—could refine our analyses even further. This approach, known as *fractionalization*, involves dividing a complex process into manageable segments,

with each segment handled by a dedicated agent. *Dominance* in this context refers to the strategic control and optimization of each segment by its respective agent, ensuring that the overall system maintains coherence and maximizes efficiency. By implementing a system based on these principles, not only is the granularity of the obtained insights improved, but the overall effectiveness and adaptability of the system within dynamic HCI environments are also enhanced. Similarly, when applied to the *Evaluation of Artworks*, this structured approach allows specialized agents to precisely manage different stages of artistic creation and interpretation, thereby enhancing the precision and depth of art evaluations. Consequently, our ArtMentor is structured around multi-agent concepts based on fractionalization and dominance principles.

2.3.2 Living Artifact and Iterative Upgrades. Following the *CoAuthor* research, Lee proposed a dynamic and adaptive framework aimed at continuously enhancing technologies for writing assistance [34]. This design space, developed through collaborations with experts from disciplines including HCI, Natural Language Processing, Information Systems, and Education, encompasses five key dimensions—*task, user, technology, interaction, and ecosystem*—and involved a comprehensive analysis of 115 papers to map the landscape of writing assistants. The framework is designed as a *living artifact*, intended to evolve through community contributions of new research, annotations, and discussions, keeping pace with advancements in the field. However, while this framework is insightful, it still relies on a traditional data collection system. Inspired by Talebirad's approach to enhancing LLMs through multi-agent systems [58], our ArtMentor employs a similar structure to address these limitations. By adopting a multi-agent architecture, we facilitate *iterative upgrades*, enabling both the dataset and the supporting system to remain dynamic and responsive to emerging needs and developments. This strategy not only enhances the adaptability of our system but also improves its capability to handle complex tasks efficiently, reflecting the collaborative environment and knowledge exchange among intelligent agents envisioned by Talebirad.

3 DESIGN PRINCIPLES FOR ArtMentor

In this section, we discuss the design of four main components of ArtMentor: *a multi-agent data collection system, an HCI dataset, a data analysis system*, and *an iterative upgrades system.* The design of ArtMentor adheres to the principles of adjusting evaluation granularity, providing immediate feedback, and progressively approaching target capabilities [2]. To design appropriate evaluation granularity, we draw inspiration from Henri Bergson's exploration of consciousness [7], shifting the focus from evaluating the physical aspects of art pieces to assessing the creative abilities of the students who create them, thereby formulating nine dimensions. Immediate feedback is linked to our approach of HCI, where MLLMs continuously engage with art teachers and students, and this feedback is meticulously recorded. Instead of directly evaluating the art pieces, we divide the process into multiple sub-processes, including entity recognition, commentary generation, and suggestion generation, thereby embodying the principle of progressively approaching target skills.

3.1 Adjusting Evaluation Granularity

Our research extends prior studies by designing a multi-agent data collection system, which serves as a core component of ArtMentor [57]. Art, as a unique form of expression, is deeply rooted in the pursuit of the essence of life, providing individuals with spiritual fulfillment [66]. In the era of big data, the evaluation of artworks faces several technical challenges, primarily manifested in four areas: data dependency, limitations on creativity, emotional reduction, and the loss of intrinsic meaning [27]. For instance, instrumental rationality often causes data to become a rigid constraint on both teachers' and students' creative interpretations of art.

To address these challenges, our system is based on the principle of relational self-expression, which elevates the evaluation process beyond mere objectivity. The core concept of relational self-expression draws inspiration from Henri Bergson's philosophy of creative evolution, emphasizing the notion that relation precedes individuality. To fully implement this idea, we incorporate insights from Bergson's exploration of consciousness [7]. Bergson's emphasis on direct experience and pre-reflective consciousness seeks to capture the undistorted reality of conscious states. Similarly, our system aims to evaluate artworks in their most authentic form. The system integrates multiple data types, focusing on images, audio, and textual data. Moreover, the evaluation process is implemented within a CHI framework, leveraging GPT-4o to assist art teachers in scoring. This hybrid approach combines machine intelligence with human expertise, ensuring both precision and contextual depth in artistic evaluation.

Our evaluation framework encompasses multiple dimensions, each designed to capture the unique characteristics of artworks. These dimensions include formative creativity (**realism**, **deformation**, **imagination**), color expressiveness (**color richness**, **color contrast**), line work richness (**line combination**, **line texture**), and conceptual thinking (**picture organization**, **transformation**). This dimensional framework was developed by a member of our research team and represents an innovative contribution to the field, though the corresponding work is yet to be formally published.

- **Realism**: Evaluates the artwork's ability to realistically reproduce subjects, capturing the precision and accuracy of representation [9].
- **Deformation**: Assesses the artwork's capacity to transform and recreate reality, reflecting artistic innovation and reinterpretation [55].
- **Imagination**: Examines the creativity and originality within the artwork, highlighting the artist's ability to introduce new and novel perspectives [53].
- **Color Richness**: Evaluates the diversity and aesthetic harmony of the color palette used in the artwork [44, 49].
- **Color Contrast**: Assesses the visual impact and vibrancy of colors, focusing on how contrasting hues interact within the composition [67].
- **Line Combination**: Evaluates the arrangement and structural coherence of lines, examining how line elements contribute to the overall form [43].
- **Line Texture**: Examines the expressiveness and tactile quality of line work, exploring the textural effects achieved through line variations [19].

- **Picture Organization**: Evaluates the overall layout and arrangement of elements within the artwork, assessing compositional balance and spatial logic [43].
- **Transformation**: Assesses the artist's ability to transform abstract concepts into tangible artistic expressions, exploring the depth of conceptual execution [20].

This multidimensional evaluation framework, grounded in the principle of relational self-expression, represents an AI-driven approach to artistic evaluation. It is based on the processes of data collection, organization, and accumulation, transitioning from object-centric, observable, and dynamic assessments to human-centered evaluations. By constructing a human-machine collaborative evaluation system, the framework ultimately aims to transcend the surface self and reach the deep self, fostering the creative and emotional expression of artistic life.

3.2 Providing Immediate Feedback

The design principles of the HCI dataset should prioritize both the evaluation outcomes and the comprehensive recording of all user-system interactions. The GOMS (Goals, Operators, Methods, and Selection Rules) model provides a systematic framework for tracking interactions, emphasizing the centrality of operators (user actions) and methods (task execution). This decomposition facilitates researchers to isolate and analyze each component, clarifying how users, such as art educators, navigate the system, refine their evaluations, and adjust feedback based on AI-assisted recommendations. [32].

Leveraging the GOMS model, which emphasizes breaking down complex tasks into smaller, manageable components [52, 56], we advocate dividing interaction processes into distinct phases to capture decision-making and feedback loops holistically. For example, in the initial phase of artwork evaluation, the dataset must capture every interaction between the educator and the *entity recognition agent*, including the system's identification results, the educator's revisions, and the final outcome. Subsequently, the *review generation agent* records initial user ratings and comments, followed by agent revisions, user modifications, and final adjustments, culminating in the user's submission. Finally, the *suggestion generation agent* data encapsulates the suggestions provided by MLLMs for improving the artwork, which can be analyzed to understand how AI-assisted feedback is utilized and acted upon in real-world educational contexts.

Each phase is systematically recorded, ensuring a comprehensive dataset for analyzing and improving MLLM capabilities. This comprehensive methodology is in line with the GOMS model's emphasis on evaluating human-system interaction by understanding not only the results, but also the cognitive steps taken throughout the process [26]. Moreover, the ability to refine each agent's performance based on the interaction data ensures that the dataset remains an integral tool in the ongoing evolution of MLLM-assisted artwork evaluation [46].

3.3 Approaching Target Capabilities

The data analysis system should enable designers to extract meaning from interactions and analyze them based on their own design goals [10]. Instead of solely depending on subjective scoring of the

procedural data to evaluate MLLMs' assistance capabilities—often prone to bias—the system facilitates a more objective, iterative evaluation process across multiple rounds. Art teachers are empowered to modify the data, prompting MLLMs to regenerate scores, reviews, and suggestions in subsequent rounds based on these adjustments. In terms of assistance capabilities, we divide them into four categories: *entity recognition*, *style evaluation*, *scoring*, and *text generation* (*reviews* and *suggestions*). These capabilities are detailed as follows:

- **Entity Recognition Capability**: Inspired by machine learning classification metrics, we define accuracy, recall, and other metrics based on the interaction between the entity recognition agent and art educators. Following each round of teacher-driven entity modifications (deleting or adding), we quantify the MLLMs' entity recognition capability, offering high interpretability through successive rounds of human-AI interaction.
- **Style Evaluation Capability**: We assess the MLLMs' style evaluation capability by measuring the extent to which art educators accept or reject the styles identified by the MLLMs, calculated by whether the teacher deletes the recognized styles.
- **Scoring Capability**: We extract the scores from the reviews to evaluate MLLMs' initial scoring of artworks, the manual scores given by teachers, and the scores regenerated by the MLLMs after human modifications. By comparing the initial MLLMs scores with teachers' scores and tracking the similarity across multiple score generation rounds, we assess the progression of MLLMs' scoring capability.
- **Text Generation Capability**: Both reviews and suggestions are treated as text outputs that can be evaluated using the same metrics. Utilizing natural language processing techniques, we divide the MLLMs' text generation capability into two main parts: the degree of modification and the similarity of the text. These metrics are derived from the modification length between initial and revised texts, and the word similarity measured through tokenization and semantic analysis over successive rounds.

4 METHODOLOGY

According to principles in Section 3, we develop ArtMentor, which includes 9 HCI processes, as shown in Figure. 2. First, the HCI dataset comprises these nine HCI processes, all of which are documented by the multi-agent data collection system and detailed in Section 4.1. Subsequently, we extract the significance of the processes into metrics, and provide feedback for iterative upgrades system through a data analysis system in Section 4.2.

4.1 Documenting Processes of the Multi-agent Data Collection System for an HCI Dataset

4.1.1 Art Teachers and E-Agent Interaction Processes. In the multi-agent data collection system, the E-Agent is responsible for extracting and classifying entities from artworks using a MLLM. We demonstrate the details of Interaction Processes.

Process 1 (P1): Entity List Recognition. The E-Agent initiates the process of recognizing entities from the given artworks. This process can be described by the following equation:

$$E_i = Ag_e(ML, Art_i, Ent\_prm, \Theta, \Gamma), \qquad (1)$$
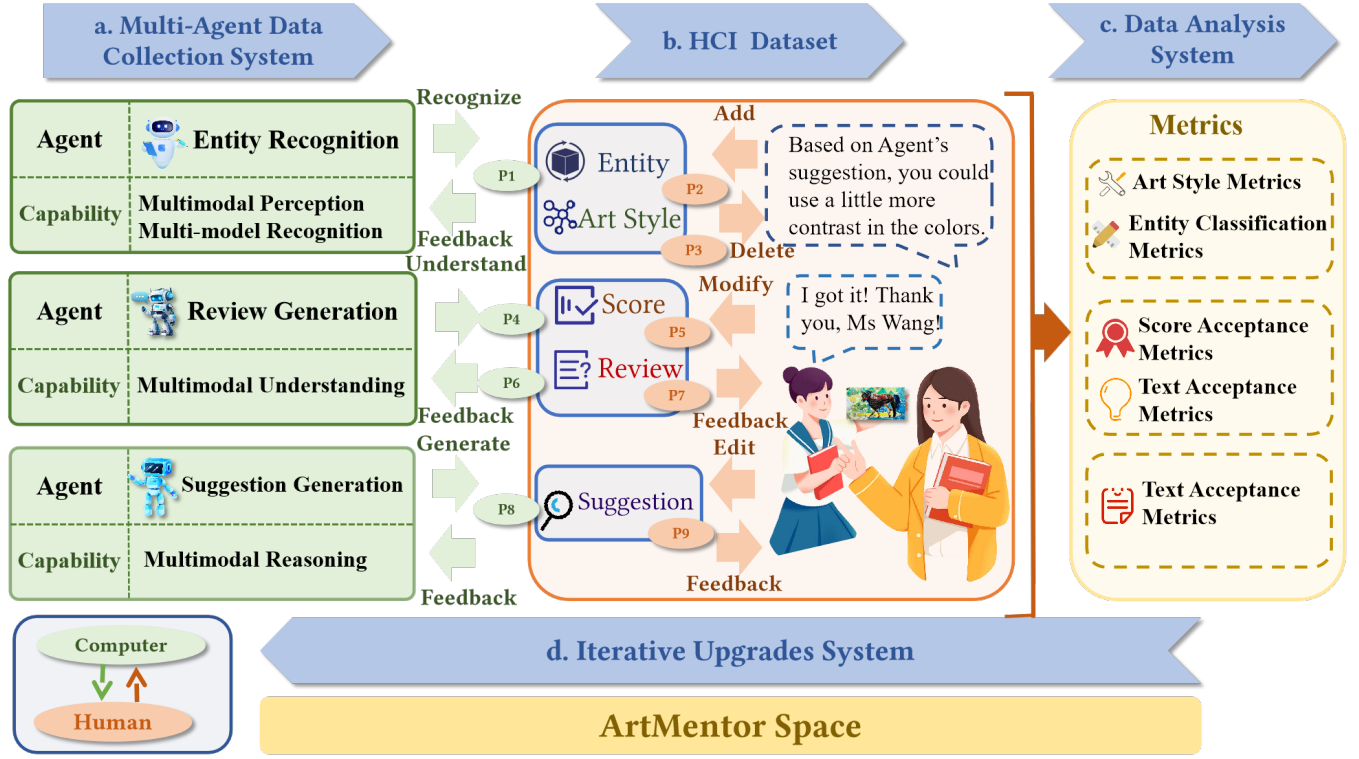
Figure 2: ArtMentor Space comprises four primary components: a. Multi-Agent Data Collection System, b. HCI Dataset, c. Data Analysis System, d. Iterative Upgrades System. The Multi-Agent Data Collection System includes three agents: Entity Recognition Agent (E-Agent), Review Generation Agent (R-Agent), and Suggestion Generation Agent (S-Agent). Both R-Agent and S-Agent perform nine dimensions, such as Realism and Deformation. Additionally, we have outlined nine HCI processes (from P1 to P9), where processes initiated by the **computer** are marked in green and those initiated by the **human** are marked in orange. After data collection by the Multi-Agent system, we obtain an HCI dataset. We then apply five metrics to evaluate these four capabilities. Based on the evaluation results, we aim to iteratively upgrade capabilities that underperform in the future.

where E_i is the set of entities recognized from the i-th artwork, Ag_e represents the E-Agent driven by the specific MLLM (ML), Art_i is the i-th artwork being analyzed, Ent_prm is the prompt provided to the MLLM to guide the entity extraction, Θ is a set of parameters used in the extraction process, and Γ is the set of dimension-related hyperparameters.

Process 2 (P2): Entity Right List Addition. During this process, the teacher adds new correct entities that are not already present in the set E_i. The set of entities newly added by the teacher is defined by the following equation:

$$R_i = \{e_{im} \mid m = 1, 2, 3, \ldots, M_i, \forall e_{im} \notin E_i\}, \qquad (2)$$

where R_i represents the set of entities added by the teacher, and $e_{im} \notin E_i$ for all m.

Process 3 (P3): Entity Wrong List Deletion. The teacher removes incorrect entities from E_i, defined as:

$$W_i = \{e_{iq} \mid q = 1, 2, 3, \ldots, Q_i, \forall e_{iq} \in E_i\}, \qquad (3)$$

where W_i is the set of entities identified as incorrect by the teacher, and $e_{iq} \in E_i$ for all q.

Final Entity List Update. Following the revisions of correct and incorrect entities, the final set of entities for each artwork is updated by incorporating the revised correct entities and excluding the incorrect ones. As defined by Eqs. 1, 2, and 3, the final entity list is expressed as:

$$\hat{E}_i = \overline{E_i} = E_i \cup R_i \setminus W_i, \qquad (4)$$

where $\hat{E}_i$ denotes the final revised list of recognized entities. The new $\overline{E_i}$ represents the updated set of entities for the i-th artwork, revised by the art teacher. After revision, $\hat{E}_i$ is equivalent to $\overline{E_i}$.

4.1.2 Art Teachers and R-Agent Interaction Processes. The R-Agent is responsible for generating reviews for artworks, which involves both scoring (ranging from 1 to 5 as integers) and textual review generation. The primary processes involved are: *review generation*, *review modification*, *score extraction*, and *score adjustment*.

Process 4 (P4): Review Generation. The R-Agent generates a review for the k-th dimension of the i-th artwork as defined by the following equation:

$$Rev_{ik} = Ag_r(ML, Art_i, Rev\_Pmt_k, \hat{E}_i, \theta_k, \gamma_k, \overline{Rev}_{ik}, f), \qquad (5)$$

where Rev_{ik} is the review generated for the k-th dimension of the i-th artwork. If $\overline{Rev}_{ik}$ is empty, the review is solely generated by the computer; if $\overline{Rev}_{ik}$ contains content, it indicates collaboration between the human and the computer in the review process. Ag_r is the Review-Agent, ML denotes the specific MLLM, Art_i refers to the artwork in question, Rev_Pmt_k is the prompt for the review, $\hat{E}_i$ is the final list of entities updated through interaction with the E-Agent and humans as per Eq. 4, θ_k and γ_k are the parameters and dimension-related hyperparameters used in generating the review, respectively, and f is the function used for extracting scores from the reviews.

Process 5 (P5): Review Modification. During this process, the teacher modifies the generated review. The modifications are quantified by the following equation:

$$\overline{Rev}_{ik} = Rev_{ik} \cup Ins_{ik} \setminus Del_{ik}, \tag{6}$$

where $\overline{Rev}_{ik}$ represents the review following human modifications. This review may be co-generated by both the computer and the human if Rev_{ik} as defined in Eq. 5 is not empty. If Rev_{ik} is empty, it indicates no computer involvement in the initial review generation. Ins_{ik} denotes the review content inserted by the teacher for the k-th dimension of the i-th artwork, and Del_{ik} denotes the review content removed from the same dimension.

Process 6 (P6): Score Extraction. Scores are extracted from the generated review as follows:

$$S_{ik} = f(Rev_{ik}), \tag{7}$$

where S_{ik} represents the score for the i-th artwork along the k-th dimension. The function f extracts the score from the review Rev_{ik}, specifically deriving from the procedures outlined in Process 4.

Process 7 (P7): Score Adjustment. This process involves the modification of scores as detailed below:

$$\overline{S}_{ik} = S_{ik} - \Delta s, \tag{8}$$

where $\overline{S}_{ik}$ denotes the adjusted score for the i-th artwork on the k-th dimension, and Δs represents the human modification, which may be either an increase or a decrease. All scores are integers within a five-point scale.

Final Review Submission. In the concluding process, scores and reviews are finalized through collaborative efforts between human and computer. The initial reviews and scores produced by the computer undergo thorough scrutiny and adjustments by the teacher, enhancing their accuracy and reliability. The final review is denoted as $\hat{Rev}_{ik}$ and includes the final score $\hat{S}_{ik}$.

4.1.3 Art Teachers and S-Agent Interaction. The S-Agent generates and revises artwork suggestions, involving processes like *Suggestion Generation*, *Modification*, and *Final Submission*. Both reviews and suggestions are text-based with similar interaction processes.

Process 8 (P8): Generation:

$$Sug_{ik} = Ag_s(ML, Art_i, \ldots, \overline{Sug}_{ik}), \tag{9}$$

where Sug_{ik} is the suggestion, Ag_s is the S-Agent, and $\overline{Sug}_{ik}$ is the co-generated suggestion.

Process 9 (P9): Modification:

$$\overline{Sug}_{ik} = Sug_{ik} \cup InsSug_{ik} \setminus DelSug_{ik}, \tag{10}$$

where $\overline{Sug}_{ik}$ is the final suggestion, $InsSug_{ik}$ is the inserted content, and $DelSug_{ik}$ is the removed content.

Final Submission: The finalized suggestions $\hat{Sug}_{ik}$, crafted through collaboration, are submitted.

4.2 Evaluation from the Data Analysis System for Iterative Upgrades System

To evaluate assistance capabilities of MLLMs, our data analysis system covers diverse metrics: *Entity Classification Metrics* (inspired by machine learning classification metrics), *Art Style Metrics*, *Score Acceptance Metrics*, and *Text Acceptance Metrics*. The insights gained from these metrics guide the iterative enhancement of capabilities.

4.2.1 Entity Classification Metrics for E-Agent. For the E-Agent, we demonstrate the design of entity classification metrics derived from the confusion matrix. We define *True Positive (TP)*, *Misrepresentation (MR)*, *False Positive (FP)*, and *False Negative (FN)* as follows:

- **TP**: The number of entities correctly identified in the artwork by the E-Agent reflecting accurately recognized entities. The calculation is expressed as:

$$TP = |E_i| - |W_i|. \tag{11}$$

- **MR**: The number of entities incorrectly identified as other entities by the E-Agent, e.g., a "horse" misidentified as a "donkey". The calculation follows:

$$MR = \min(|W_i|, |R_i|). \tag{12}$$

- **FP**: The number of non-existent entities recognized by the E-Agent, e.g., mistakenly identifying nonexistent water in an artwork. The calculation is adjusted for misrepresentations:

$$FP = \max(|W_i| - MR, 0). \tag{13}$$

- **FN**: The number of entities overlooked by the E-Agent. This involves entities present in the artwork but not recognized, calculated by:

$$FN = \max(|R_i| - MR, 0). \tag{14}$$

Based on these definitions, we redefine the metrics of *accuracy*, *precision*, *recall*, and *F1-score* to evaluate the E-Agent's entity recognition capabilities:

$$
\begin{aligned}
Accuracy &= \frac{TP}{TP + FP + FN + MR}, \\
Precision &= \frac{TP}{TP + FP + MR}, \\
Recall &= \frac{TP}{TP + FN + MR}, \\
F1 &= \frac{2 \times Precision \times Recall}{Precision + Recall}.
\end{aligned}
\tag{15}
$$

These calculations provide a comprehensive evaluation of MLLMs' performance in entity recognition within artworks, facilitating targeted improvements to the model.

4.2.2 Art Style Metrics for E-Agent. We conceptualize art style as a distinct entity and introduce the **Art Style Sensitivity (ASS)** metric to assess the E-Agent's capability to accurately recognize and evaluate various art styles. The metric is defined as follows:

$$ASS = 1 - \frac{D}{N}, \tag{16}$$

where N represents the total number of art styles identified, and D denotes the number of incorrect recognitions as flagged and corrected by the art teacher.

4.2.3 Score Acceptance Metrics for R-Agent. As the scoring of artworks becomes increasingly prevalent in automated assessment systems, it is crucial to evaluate the scoring capabilities of the R-Agent. We focus on three key aspects: First: the *score difference* between **initial** scores provided by MLLMs and the scores assigned by art teachers in subsequent rounds (*Process 4* to *Process 5*). Second, the *consistency of scoring* between MLLMs and art teachers across all rounds. Third, the *volatility of scores* either from art teachers or MLLMs.

Score Difference (SD): This metric quantifies the deviation between the initial MLLM-generated scores and the modified scores assigned by art teachers in subsequent rounds (*Process 5*). The equation is defined as:

$$\text{SD} = \frac{1}{N} \sum_{i=1}^{N} \left| S_{\text{MLLM}}^{(1)} - S_{\text{art teacher}}^{(i)} \right|, \tag{17}$$

where $S_{\text{MLLM}}^{(1)}$ represents the initial MLLM score, $S_{\text{art teacher}}^{(i)}$ denotes the art teacher's modified score in the i-th round, and N is the total number of rounds.

Score Consistency (SC): This metric evaluates the alignment between art teachers' scores and those of MLLMs by measuring the Spearman correlation coefficient (ρ) between the scores given by MLLMs and the scores modified by art teachers. The Spearman coefficient is a non-parametric measure of rank correlation that assesses the monotonic relationship between two variables. It is calculated as follows:

$$\text{SC} = \rho(S_{\text{MLLM}}, S_{\text{art teacher}}), \tag{18}$$

where S_{MLLM} and $S_{\text{art teacher}}$ are vectors representing the series of scores from MLLMs and the corresponding modified scores by art teachers, across all rounds.

Score Volatility (SV): This metric quantifies the stability of scoring by either art teachers or MLLMs across all rounds, reflecting the consistency of the scoring process. It is calculated using the standard deviation:

$$\text{SV} = \text{std}(S_{\text{scorer}}), \tag{19}$$

where S_{scorer} represents the vector of scores modified by a specific scorer, either an art teacher or an MLLM. This metric is crucial for identifying instances of erratic scoring behavior that could undermine the reliability of the analysis.

4.2.4 Text Acceptance Metrics for R-Agent and S-Agent. For R-Agent and S-Agent, natural language generation represents a core capability, producing both reviews and suggestions in textual form. However, analyzing text within HCI contexts can be challenging. To address this, we propose evaluating text through two key metrics: text length and word similarity. These metrics respectively provide insights into the conciseness and relevance of the generated text.

Text Modification Rate (TMR): This metric quantifies the acceptance of MLLM-generated text (e.g., reviews, suggestions) by calculating the ratio of characters modified by the art teacher to

the original text. The formula is defined as:

$$\text{TMR} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{len}(T_{\text{MLLM}}^{(i)}) - \text{len}(T_{\text{removed}}^{(i)})}{\text{len}(T_{\text{MLLM}}^{(i)}) + \text{len}(T_{\text{added}}^{(i)})}, \tag{20}$$

where $T_{\text{added}}^{(i)}$ and $T_{\text{removed}}^{(i)}$ represent the number of characters added and removed by the art teacher in the i-th round, respectively, and $T_{\text{MLLM}}^{(i)}$ is the original number of characters in the MLLM-generated text.

Text Similarity (TS): We assess the similarity between the original MLLM-generated text and the modified text using cosine similarity metrics. Prior to similarity computation, both texts are vectorized using the Bag of Words (BoW) model, referred to as "Wordbag" in our context. This model represents each text as a vector, where each dimension corresponds to a word in the vocabulary. The value in each dimension represents the frequency of the corresponding word in the text. The vectorization process is detailed as follows:

$$T_{\text{MLLM}} = \text{Wordbag}(T_{\text{MLLM}}),$$
$$T_{\text{art teacher}} = \text{Wordbag}(T_{\text{art teacher}}), \tag{21}$$

where T_{MLLM} and $T_{\text{art teacher}}$ are the vectorized representations of the MLLM-generated and art teacher-modified texts, respectively.

Cosine similarity is then used to compute the degree of similarity:

$$\text{TS} = \text{Cosine}(T_{\text{MLLM}}, T_{\text{art teacher}}) = \frac{T_{\text{MLLM}} \cdot T_{\text{art teacher}}}{\|T_{\text{MLLM}}\| \|T_{\text{art teacher}}\|}. \tag{22}$$

5 GUIDE FOR ArtMentor SPACE

5.1 Data Collection Interface

The user interface and interaction design of the ArtMentor space are specifically developed to provide art teachers with an intuitive and structured tool for evaluating artworks. The primary objective of this interface is to support detailed art analysis through a series of structured interaction processes.
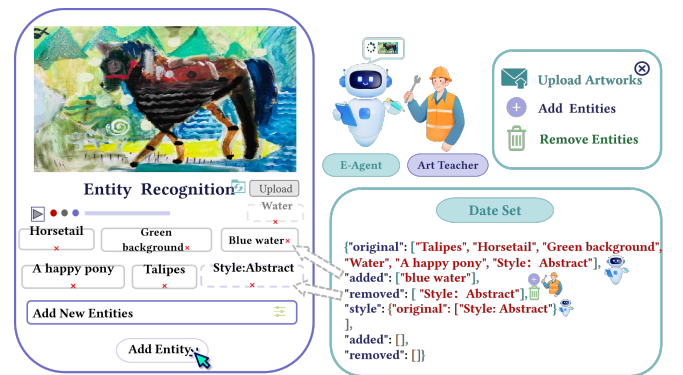
Figure 3: E-Agent and art teacher interaction collection.

As shown in Figure. 3, the interface allows art teachers to upload an artwork for analysis and provides multiple evaluation sections based on different artistic dimensions. After uploading the artwork, the interface displays the artwork and automatically recognizes

and lists entities within the artwork. Art teachers can add or remove entities to optimize the recognition results. Finally, the HCI processes saved in JSON files.

As shown in Figure. 4, the interface offers a set of categorized evaluation tools. Art teachers can select different dimensions (e.g., realism, deformation) to generate or manually input scores and reviews. Each dimension has specific evaluation tools, including buttons for generating review and suggestions from MLLMs.
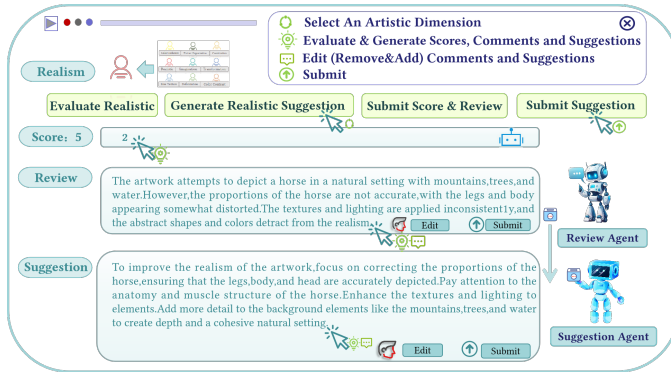

Figure 4: R & S-Agent and art teacher interaction collection.

5.2 Data Collection Procedure

5.2.1 Participants and Artworks. In the ARTMENTOR space, we recruited five art teachers from diverse educational backgrounds, including both pre-service and in-service art teachers. The selection of participants was aimed at offering a range of perspectives to facilitate a comprehensive evaluation of elementary students' artworks. The artworks used in this study were categorized into three groups: narrative illustrations (1-3), Chinese ink paintings (4-7), and artworks following the Egyptian frontal law (8-20), with each number range corresponding to its respective category.

5.2.2 Art Evaluation Details. Participants were provided with detailed instructions on how to interact with the ARTMENTOR space. They were required to evaluate each piece of artwork across multiple dimensions, spending at least ten minutes on each evaluation. To enhance the multidimensional nature of the evaluation process and capture more subtle aspects of artistic expression, students were allowed to provide audio explanations of their artworks. These audio explanations allowed art teachers to gain insights into the students' creative processes, intentions, and emotional connections to their work, aligning with our principle of evaluation beyond dehumanization. The guidelines emphasized three primary requirements: (1) ensuring that each evaluation is thorough, covering multiple artistic dimensions; (2) actively using the MLLM-generated reviews and suggestions provided by the system to enhance their evaluations; (3) carefully considering the students' audio explanations to capture nuanced information about the artistic process and intent.

5.2.3 Post-Evaluation Interviews. After completing each evaluation session, participants were asked to participate in an interview designed to gather feedback on their experience with the ARTMENTOR space. The interview focused on five key areas: participants'

perceptions of the accuracy of the MLLM's entity recognition, the usability and user interaction experience with the space, the effectiveness of MLLM in assisting with scoring and review or suggestion generation, the MLLM's performance across different functional dimensions, and the relevance of the MLLM's suggestions in supporting elementary students' artistic development. The detailed interview questionnaire is provided in Appendix A. The content of the interviews will be used to support the result analysis, and some materials can be accessed at link.

5.3 HCI Dataset Overview

GPT-4o [1] has been selected to assist in art education due to its advanced capabilities in MLLMs. The dataset comprises evaluative feedback from five art teachers on a diverse collection of 20 artworks, covering multiple key dimensions and providing a rich resource for analysis. It is worth noting that the teachers and students who contributed to this dataset are from the first and second grades of a primary school, offering a valuable educational perspective from early childhood education.

The dataset includes 20 JSON files documenting the GPT-4o's entity and style recognition results for each artwork. Additionally, there are 360 JSON files that record detailed reviews and suggestions provided by the teachers for each dimension. This extensive dataset, which chronicles 380 sessions, lays a solid foundation for studying the application of MLLMs in art education.

6 RESULTS ANALYSIS SYSTEM IN ARTMENTOR

In this section, we regard art style as a distinct entity. Consequently, we analyze the results concerning entity classification metrics and art style metrics in Section 6.1. Next, we examine the score generation capabilities of GPT-4o in Section 6.2. Finally, we evaluate the review and suggestion capabilities of GPT-4o in Section 6.3.

6.1 Analysis of Entity Recognition Capability

To evaluate the entity recognition capabilities of GPT-4o, we **initially** focus on *art style recognition* as it offers a comprehensive insight into the artwork, forming the basis of our analysis. **Subsequently**, we examine the *average recognition capabilities* across various entities. **Finally**, we delve into the finer details within the artwork to assess *nuanced recognition performance*.

6.1.1 How capable is GPT-4o in recognizing art styles? As depicted in Figure. 5, GPT-4o demonstrates a robust art style recognition capability with an 80% Art Style Sensitivity (ASS), as detailed in Section 4.2.2. Specifically, it accurately identifies narrative illustrations (artworks 1-3) and adheres to the Egyptian frontal law (artworks 8-20).

However, it occasionally misidentifies Chinese ink paintings (artworks 4-7) as watercolors. This error likely arises from a misunderstanding by GPT-4o that Chinese ink paintings can indeed include a variety of colors, not just simply black and white. This misconception underscores the need for model optimization to better understand and differentiate between complex art styles, paving the way for future enhancements.

6.1.2 What is the average entity recognition capability of GPT-4o? As depicted in Figure. 7, precision and F1-score are notably high.
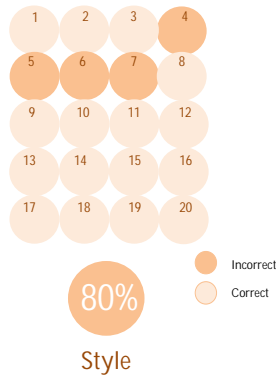
Figure 5: Recognition of art styles by GPT-4o across 20 artworks (Artwork Numbers 1-20).

The elevated precision may result from a high rate of false negatives (FN), as explained by the precision equation (Eq. 15) and the FN equation (Eq. 14). A detailed examination reveals that $|R_i|$ exceeds $|W_i|$, suggesting art teachers often add more entities than they remove, indicative of a potential bias towards over-recognition. This bias ensures no significant entities are overlooked, aligning with GPT-4o's strength in minimizing false positives—a key factor in applications that demand high predictive accuracy. Moreover, the F1-score of 0.881 demonstrates a strong, balanced performance in entity recognition between precision and recall.

Interestingly, accuracy (0.833) and recall (0.836) closely align, suggesting minimal false positives (FP), as indicated by the precision formula (Eq. 15) and the false positive formula (Eq. 13). The near equivalence of $|W_i|$ and misrepresentations (MR) supports high precision, showing that art teachers typically replace entities with precision. This behavior underscores the model's adaptability and conservative approach to entity handling, enhancing its generalization across various art styles.

6.1.3 How does GPT-4o perform in recognizing artwork details? To clarify the results further, we illustrate them with a diagram of four waffle charts as shown in Figure. 6.

The **first** observation highlights a significant discrepancy between recall (see Figure. 6c) and accuracy (see Figure. 6a) in the thirteenth artwork. Unlike other artworks where these metrics are usually similar, this piece features many abstract figures, potentially causing a high number of false positives. The **second** observation concerns the near absence of false positives. In most artworks, precision is exceptionally high (often 1.000), as shown in Figure. 6b, suggesting that the model rarely misidentifies non-existent entities and maintains stable performance across various artworks. However, styles like Chinese ink painting may require further optimization to enhance recall. **Finally**, metrics indicate underperformance in the sixth and seventh artworks, classified under Chinese ink painting. Their abstract nature poses challenges for the model in accurately recognizing and classifying entities.

6.2 Analysis of Score Generation Capability

First, we examine the score volatility of both GPT-4o and art teachers to validate the reliability of score generation. Next, we investigate any significant differences between the initial scores provided by GPT-4o and those assigned by art teachers prior to HCI. Finally, we assess whether the scores from GPT-4o and art teachers converge or diverge after HCI, exploring potential changes towards consensus or further discrepancy.

6.2.1 What is the score volatility for GPT-4o and art teachers? The analysis of Score Volatility (SV) provides significant insights into the stability of scoring by both human scorers and GPT-4o prior to HCI. Art teachers demonstrate consistent scoring patterns, assigning identical scores to the same artwork across similar dimensions throughout the evaluation process. Similarly, GPT-4o, operating at a temperature setting of zero, consistently produces the same outputs for identical artworks under the same conditions. The minimal score volatility observed reinforces the reliability of the dataset, leading us to accept it as robust for further analysis.

6.2.2 What are the differences between initial scores from GPT-4o and art teachers? The Score Difference (SD) metric uncovers scoring discrepancies across various artistic dimensions. As shown in Figure. 8a, **Realism** and **Transformation** display higher SDs of 0.3208 and 0.2941, respectively, highlighting significant divergences between the model's initial assessments and human evaluations.

In contrast, categories such as **Deformation**, **Color Richness**, **Color Contrast**, **Line Combination**, **Line Texture**, and **Picture Organization** show minimal SDs, indicating a strong concordance between GPT-4o outputs and art teacher adjustments.

Imagination, with a moderate SD of 0.5000, signals the potential for further model tuning to achieve closer alignment with expert judgments. Overall, while GPT-4o generally agrees with human scoring, it requires targeted improvements in areas like **Imagination** to *refine its evaluative accuracy*.

6.2.3 How do scores from GPT-4o and art teachers change after human-computer interaction? The Score Consistency (SC) analysis, employing the Spearman correlation coefficient, reveals strong alignment between GPT-4o and art teachers' scores post-HCI, as depicted in Figure. 8b. SC values for key dimensions are: **Realism** (0.9438), **Deformation** (0.9655), **Imagination** (0.5209), and **Transformation** (0.6555), among others. Notably, **Realism** saw significant score convergence in later rounds, *suggesting effective model-human integration through iterative feedback*.

However, lower SC values in **Imagination** and **Transformation** highlight areas needing further model calibration to better match human assessments. Overall, the post-HCI data indicates a strong general agreement, with ongoing model refinements crucial for uniform scoring accuracy across all dimensions.

6.3 Analysis of Review and Suggestion Generation Capability

We evaluate GPT-4o's ability to generate reviews and suggestions using two metrics: the Text Modification Rate (TMR) and Text Similarity (TS). These metrics quantify how well the generated texts

(a) Accuracy.

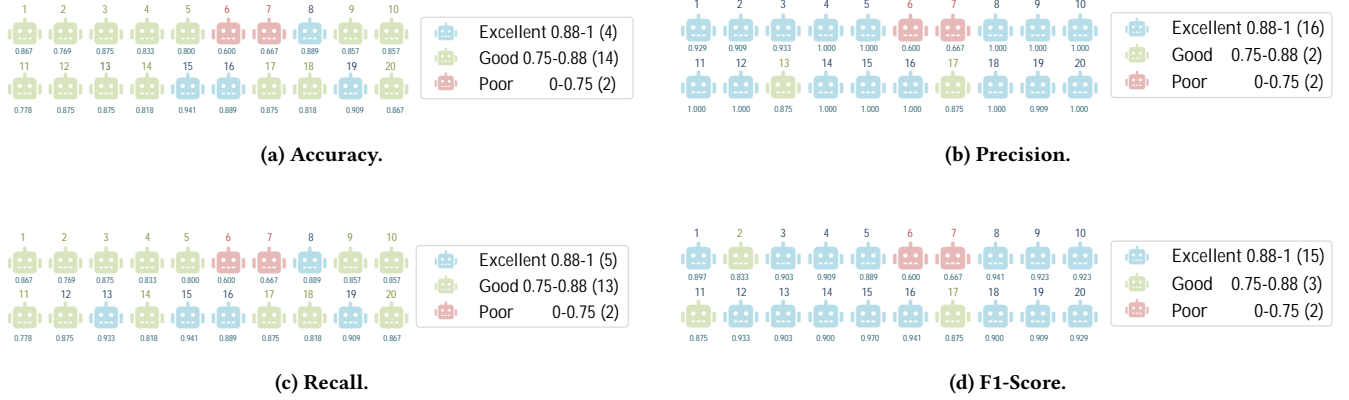(b) Precision.

(c) Recall.

(d) F1-Score.

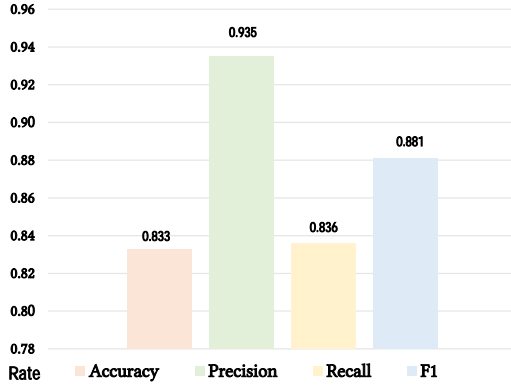Figure 6: Entity classification metrics for GPT-4o across 20 artworks (Artwork Numbers 1-20).

Figure 7: Average entity recognition capability of GPT-4o.

meet art teachers' expectations and modifications, providing insights into the effectiveness of the model's text generation. Detailed analyses in the following subsubsections will explore the distinct capabilities of GPT-4o in generating reviews and suggestions.

6.3.1 How Effective is the Review Generation Capability of GPT-4o? The Text Modification Rate (TMR) and Text Similarity (TS) metrics provide insights into the alignment of GPT-4o-generated reviews with expert evaluations.

As illustrated in Figure. 9, the **TMR** values, which indicate the extent of textual modifications by art teachers, are shown for various dimensions. These include: **Realism** (0.881), **Deformation** (0.882), **Imagination** (0.875), **Color Richness** (0.971), **Color Contrast** (0.969), **Line Combination** (0.957), **Line Texture** (0.978), **Picture Organization** (0.964), and **Transformation** (0.956). These metrics suggest that the generated reviews were closely aligned with the expectations, requiring only minimal adjustments.

Additionally, as depicted in Figure. 10, the **TS** metrics highlight the semantic consistency post-modification, with high similarity scores noted across various dimensions: **Realism** (0.992), **Deformation** (0.997), **Imagination** (0.978), **Color Richness** (0.995), **Color Contrast** (0.995), **Line Combination** (0.998), **Line Texture** (0.986),

Picture Organization (0.993), and **Transformation** (0.999). These scores **confirm** that *the modifications by art teachers **maintained** the core content and intent of the original reviews*, verifying GPT-4o's capability to generate contextually appropriate and stylistically precise content.

6.3.2 How Effective is the Suggestion Generation Capability of GPT-4o? The effectiveness of GPT-4o in generating suggestions is also assessed using the **TMR** and **TS** metrics.

As shown in Figure. 11, the **TMR** values, based on the length of added and deleted content, indicate that GPT-4o performs relatively well in dimensions like **Imagination** (0.968), **Realism** (0.940), and **Deformation** (0.895), where fewer modifications were necessary. In contrast, lower TMR values in **Transformation** (0.719), **Line Combination** (0.730), and **Line Texture** (0.731) suggest that suggestions in these areas required more extensive revisions by art teachers, both in terms of content addition and deletion.

In Figure. 12, the **TS** values, which focus on word semantics, show high similarity in dimensions such as **Realism** (0.999), **Deformation** (0.999), and **Imagination** (0.997), indicating that the core meaning of the suggestions remained largely intact. However, dimensions with lower TMR, such as **Transformation** (0.788), **Line Texture** (0.863), and **Line Combination** (0.879), also exhibit lower TS values, suggesting that not only was sentence length altered, but significant semantic changes were made as well. This pattern contrasts with review generation, where modifications affected length without drastically changing semantics, indicating that suggestion generation is less effective than review generation.

Overall, the *dimension requiring the most improvement* is **Transformation**, as it shows substantial modifications in both TMR and TS, reflecting changes in both content length and semantics. Enhancing GPT-4o's suggestion generation in this area would greatly improve its alignment with art teacher expectations.
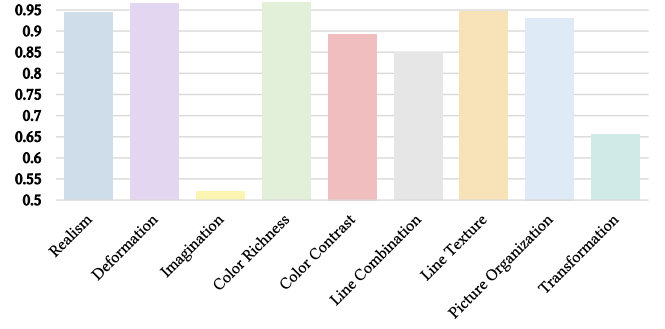
7 DISCUSSION

In this section, we delve into the intriguing question of whether an entity at the intersection of the arts, education, and AI/MLLMs
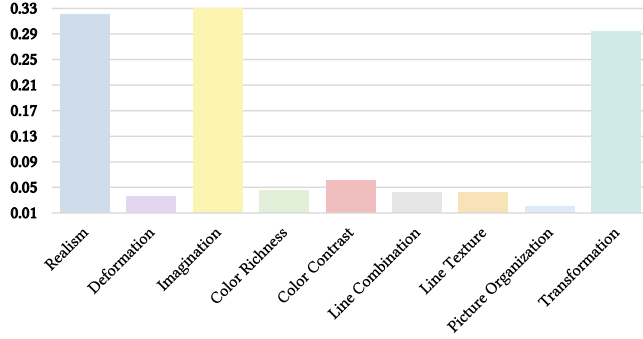
(a) Score Difference (SD).

(b) Score Consistency (SC).
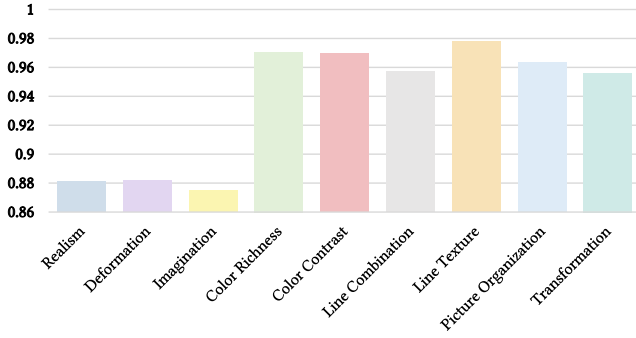
Figure 8: Score Acceptance Metrics.

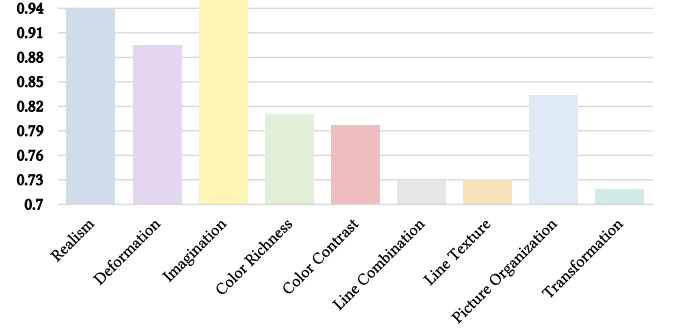Figure 9: Text Modification Rate (TMR) for R-Agent.
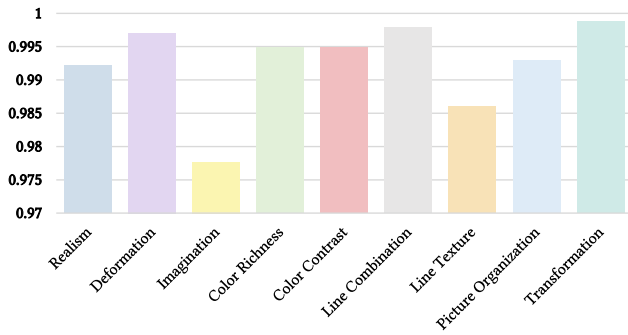

Figure 11: Text Modification Rate (TMR) for S-Agent.
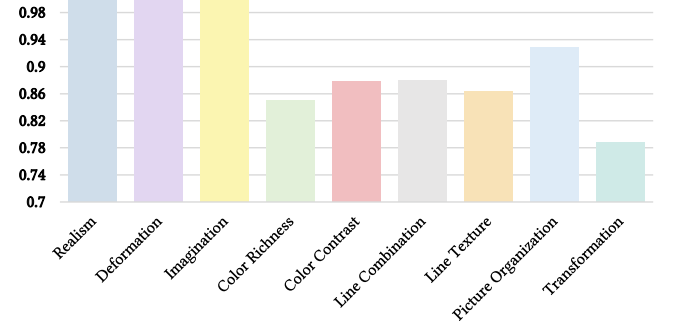

Figure 10: Text Similarity (TS) for R-Agent.

Figure 12: Text Similarity (TS) for S-Agent.

7.1 Multimodal Perception and Recognition Capabilities in GPT-4o's Entity Recognition

In our analysis of entity recognition performance, we observed two key challenges. First, art teachers tended to add more entities than they removed, suggesting that GPT-4o's initial entity recognition often overlooks relevant details. Second, GPT-4o emphasizes local features, such as facial details or limbs, rather than recognizing complete entities, leading to an over-granulation effect. This reveals a limitation in GPT-4o's ability to holistically interpret artwork.

can emerge as an independent assistant within the teacher-student-machine triadic dialogue system for the evaluation of artworks. To address this, we will systematically examine the manifestations of capabilities in multimodal perception, recognition, understanding, and reasoning across the three intelligent agents, thereby shedding light on the potential of such a MLLM to autonomously contribute to artistic assessment.

However, despite these challenges, GPT-4o demonstrates remarkable multimodal perception and recognition capabilities. The high precision and F1-score across various artworks indicate a strong ability to identify and classify entities accurately. The model's performance in recognizing art styles, as evidenced by an 80% Art Style Sensitivity (ASS), further underscores its robustness in understanding and differentiating complex artistic expressions. GPT-4o's ability to balance between holistic interpretation and detail orientation is noteworthy. While it sometimes misses broader entities, its focus on local features ensures that intricate details are not overlooked. This dual approach is particularly valuable in the context of art analysis, where both the overall style and the finer details contribute to the understanding of the artwork. The model's adaptability across different art styles, from narrative illustrations to Chinese ink paintings, highlights its generalization capabilities. Although there are areas for improvement, such as better differentiation between Chinese ink paintings and watercolors, the overall performance is commendable. This adaptability is crucial for applications that require consistent performance across a diverse range of artistic styles. The near absence of false positives is a significant strength of GPT-4o. High precision values indicate that the model is conservative in its entity recognition, avoiding the misidentification of non-existent entities. This approach is particularly important in scenarios where accuracy is paramount. While GPT-4o shows excellent performance in both entity and style recognition, there is room for further refinement. Future work could focus on enhancing the model's ability to recognize broader entities and reducing the over-granulation effect. Additionally, optimizing the model's understanding of specific art styles, such as Chinese ink painting, could further improve its performance.

In summary, GPT-4o's entity recognition and style recognition capabilities are excellent. The model's high precision, strong F1-score, and robust art style sensitivity demonstrate its effectiveness in analyzing artwork. Despite some limitations, the model's overall performance is impressive, and it holds great potential for future advancements in multimodal perception and recognition.

7.2 Multimodal Understanding Capability in GPT-4o's Review Generation

In the **Realism** dimension, we observed an encouraging trend toward convergence between GPT-4o and art teachers during the scoring process. Initial discrepancies in assessments gradually diminished as both sides adjusted through repeated interactions, ultimately leading to closer alignment. This suggests potential for improving the model's adaptability over time.

The convergence observed in the **Realism** dimension is a testament to GPT-4o's ability to learn and adapt through human-computer interaction. This adaptability is crucial for the model's long-term effectiveness, as it indicates a capacity for continuous improvement in understanding and evaluating artistic nuances. GPT-4o's performance in generating reviews that align closely with art teachers' expectations across various dimensions, as evidenced by high TMR and TS metrics, demonstrates its holistic understanding of artwork. The model's ability to maintain the core content and intent of reviews even after modifications by art teachers highlights

its robustness in generating contextually appropriate and stylistically precise content. While the overall performance is strong, a closer look at specific dimensions reveals areas for further refinement. For instance, the **Imagination** dimension shows moderate TMR and TS values, indicating a need for enhanced understanding of abstract and creative aspects of artwork. This suggests that GPT-4o could benefit from additional training data or algorithmic adjustments to better capture the subtleties of imaginative expressions. The iterative feedback loop established between GPT-4o and art teachers plays a pivotal role in refining the model's review generation capabilities. This loop enables the model to receive targeted feedback, making incremental improvements with each iteration. Continued engagement in this feedback loop is essential for achieving even greater accuracy and alignment with human evaluations. Future work should focus on expanding GPT-4o's training to include a wider range of artistic styles and dimensions, particularly those that currently show lower convergence rates. Additionally, exploring advanced multimodal understanding techniques, such as integrating visual and textual data more seamlessly, could further enhance the model's review generation capabilities.

In summary, GPT-4o's multimodal understanding capability in review generation is robust, with strong performance across various artistic dimensions. The model's ability to converge with human assessments over time, coupled with its adaptability and holistic review generation, underscores its potential as a valuable tool in art evaluation. Ongoing refinements and iterative feedback will be key to unlocking even greater capabilities in the future.

7.3 Multimodal Reasoning Capability in GPT-4o's Suggestion Generation

The analysis of GPT-4o's suggestion generation capability reveals insights into its multimodal reasoning abilities, particularly in how it integrates visual and textual information to provide constructive feedback.

GPT-4o demonstrates strong performance in certain dimensions, as evidenced by high TMR and TS values in **Imagination**, **Realism**, and **Deformation**. These high values indicate that the suggestions generated in these areas were closely aligned with art teachers' expectations, requiring minimal modifications. This suggests that GPT-4o effectively captures and reasons about the key elements of these artistic dimensions. However, the model faces challenges in more complex dimensions such as **Transformation**, **Line Combination**, and **Line Texture**. The lower TMR and TS values in these areas indicate that the suggestions often required significant revisions, both in terms of content length and semantics. This suggests that GPT-4o's multimodal reasoning capabilities may need further refinement to better understand and articulate the nuances of these dimensions. While GPT-4o shows a strong ability to holistically understand artwork, as seen in its review generation performance, the challenge lies in translating this understanding into detailed, actionable suggestions. The discrepancy between review and suggestion generation effectiveness highlights the need for the model to not only comprehend the artwork but also to provide precise and contextually relevant advice. The iterative feedback loop with art teachers plays a crucial role in enhancing GPT-4o's suggestion

generation capabilities. By receiving detailed feedback on its suggestions, the model can learn to better align its outputs with human expectations. Continued engagement in this feedback loop is essential for refining the model's multimodal reasoning abilities. To address the challenges identified, future work could explore advanced multimodal reasoning techniques. For instance, integrating more sophisticated visual analysis tools could help the model better understand complex visual elements. Additionally, leveraging techniques such as attention mechanisms and fine-grained feature extraction could enhance the model's ability to provide detailed and accurate suggestions.

In summary, GPT-4o's multimodal reasoning capability in suggestion generation shows promise, with strong performance in certain dimensions. However, there is room for improvement, particularly in handling more complex artistic dimensions. Ongoing refinements, iterative feedback, and technological advancements will be key to enhancing the model's ability to generate contextually appropriate and actionable suggestions, thereby making it an even more valuable tool in the artistic domain.

8 Conclusion

In our work, we delve into the fascinating question of whether an entity at the intersection of the arts, education, and AI/MLLMs can emerge as an independent assistant within the teacher-student-machine triadic dialogue system for artwork evaluation. To explore this, we have adopted a Human-Computer Interaction (HCI) space design and analysis approach. We have developed the ARTMENTOR space, which comprises four core components: **a. Multi-Agent Data Collection System**, **b. HCI Dataset**, **c. Data Analysis System**, and **d. Iterative Upgrades System**. The HCI dataset encompasses 380 sessions across nine dimensions of artwork evaluation, utilizing process-based data to mitigate the inherent manipulation risks associated with outcome-based data. Within the Data Analysis System, we have applied machine learning and natural language processing techniques to imbue the process data with meaning, extracting metrics that objectively reflect MLLM performance while also ensuring interpretability.

Our comprehensive exploration reveals that an entity at the intersection of the arts, education, and AI/MLLMs can indeed emerge as an independent assistant (GPT-4o) within the teacher-student-machine triadic dialogue system for the evaluation of artworks. Despite certain limitations, such as a tendency to overlook broader entities and an over-granulation effect in entity recognition, GPT-4o has demonstrated remarkable capabilities in multimodal perception, recognition, understanding, and reasoning. GPT-4o's high precision, strong F1-score, and robust art style sensitivity in entity recognition underscore its effectiveness in analyzing artwork. Its ability to adapt and converge with human assessments over time, as evidenced in review generation, highlights its potential for continuous improvement. Furthermore, the model's performance in generating contextually appropriate and stylistically precise content demonstrates its holistic understanding of artwork. While there are areas for further refinement, particularly in handling more complex artistic dimensions and reducing the over-granulation effect, the overall performance of GPT-4o is impressive. Its adaptability across different art styles and its conservative approach to entity

recognition, with a near absence of false positives, are significant strengths. In suggestion generation, GPT-4o shows promise, with strong performance in certain dimensions. However, the challenge lies in translating its holistic understanding into detailed, actionable suggestions.

In summary, GPT-4o holds great potential as a valuable tool in art evaluation within the teacher-student-machine triadic dialogue system. Ongoing refinements, iterative feedback, and technological advancements will be key to unlocking even greater capabilities in the future, thereby further solidifying its role as an independent assistant in the artistic domain.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390* (2023).

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966* 3, 1 (2023).

[5] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890* (2023).

[6] Siwar Bengamra, Olfa Mzoughi, André Bigand, and Ezzeddine Zagrouba. 2024. A comprehensive survey on object detection in Visual Art: taxonomy and challenge. *Multimedia Tools and Applications* 83, 5 (2024), 14637–14670.

[7] Henri Bergson. 1911. *Essai sur les données immédiates de la conscience.* F. Alcan.

[8] Anjanava Biswas and Wrick Talukdar. 2024. Robustness of Structured Data Extraction from In-Plane Rotated Documents Using Multi-Modal Large Language Models (LLM). *Journal of Artificial Intelligence Research* (2024).

[9] Moinak Biswas. 2021. Realism. *BioScope: South Asian Screen Studies* 12, 1-2 (2021), 158–161.

[10] Ann E Blandford, Philip J Barnard, and Michael D Harrison. 1995. Using Interaction Framework to guide the design of interactive systems. *International journal of human-computer studies* 43, 1 (1995), 101–130.

[11] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.

[12] Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Ziyong Feng, Yongle Zhao, and Yin Xie. 2024. Plug-and-play grounding of reasoning in multimodal large language models. *arXiv preprint arXiv:2403.19322* (2024).

[13] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. 2024. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. *arXiv preprint arXiv:2408.03361* (2024).

[14] Shih-Yeh Chen, Pei-Hsuan Lin, and Wei-Che Chien. 2022. Children's digital art ability training system based on ai-assisted learning: a case study of drawing color perception. *Frontiers in psychology* 13 (2022), 823078.

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

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

[17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.

[18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]

[19] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* 44, 5 (2020), 2567–2581.

[20] Yingbei Du. 2020. Research on the transformation and innovation of visual art design form based on digital fusion technology. *Applied Mathematics and Nonlinear Sciences* (2020).

[21] David W Eccles and Güler Arsal. 2017. The think aloud method: what is it and how do I use it? *Qualitative Research in Sport, Exercise and Health* 9, 4 (2017), 514–531.

[22] K Anders Ericsson. 2017. Protocol analysis. *A companion to cognitive science* (2017), 425–432.

[23] Dejan Grba. 2022. Deep else: A critical framework for ai art. *Digital* 2, 1 (2022), 1–32.

[24] Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. *arXiv preprint arXiv:2408.15769* (2024).

[25] Olga M Hubard. 2010. Three modes of dialogue about works of art. *Art Education* 63, 3 (2010), 40–45.

[26] Christian P Janssen and Duncan P Brumby. 2015. Strategic adaptation to task characteristics, incentives, and individual differences in dual-tasking. *PloS one* 10, 7 (2015), e0130009.

[27] Jing Jin and Runzhou Li. 2024. Implications, Concerns, and Transcendence of AI-Based Art Evaluation: Focusing on "Elementary School Art". *Curriculum, Teaching Material, and Method* 44, 05 (2024), 138–143. https://doi.org/10.19877/j.cnki.kcjcjf.2024.05.020

[28] Bonnie E John and David E Kieras. 1996. Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction (TOCHI)* 3, 4 (1996), 287–319.

[29] David Kadish, Sebastian Risi, and Anders Sundnes Løvlie. 2021. Improving object detection in art images using only style transfer. In *2021 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[30] Margaret Kelaher, Naomi Berman, David Dunt, Victoria Johnson, Steve Curry, and Lindy Joubert. 2014. Evaluating community outcomes of participation in community arts: A case for civic dialogue. *Journal of Sociology* 50, 2 (2014), 132–149.

[31] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 119–132.

[32] Maurice Lamb, Rachel W Kallen, Steven J Harrison, Mario Di Bernardo, Ali Minai, and Michael J Richardson. 2017. To pass or not to pass: Modeling the movement and affordance dynamics of a pick and place task. *Frontiers in psychology* 8 (2017), 1061.

[33] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5, 1 (2018), 1–10.

[34] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–35.

[35] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[36] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal Large Language Models. (2023). arXiv:2307.16125 [cs.CV]

[37] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13299–13308.

[38] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890* (2023).

[39] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-Understand: Leveraging Visual Prompts to Enable MLLMs to Comprehend What You Want. *arXiv preprint arXiv:2403.20271* (2024).

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[41] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281* (2023).

[42] Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, et al. 2024. II-Bench: An Image Implication Understanding Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2406.05862* (2024).

[43] Paul J Locher, Pieter Jan Stappers, and Kees Overbeeke. 1999. An empirical evaluation of the visual rightness theory of pictorial composition. *Acta psychologica* 103, 3 (1999), 261–280.

[44] Peng Lu, Zhijie Kuang, Xujun Peng, and Ruifan Li. 2015. Discovering harmony: A hierarchical colour harmony model for aesthetics assessment. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part III 12*. Springer, 452–467.

[45] Silverio Martínez-Fernández, Xavier Franch, Andreas Jedlitschka, Marc Oriol, and Adam Trendowicz. 2021. Developing and operating artificial intelligence models in trustworthy autonomous systems. In *International Conference on Research Challenges in Information Science*. Springer, 221–229.

[46] Heramb Nemlekar, Jignesh Modi, Satyandra K Gupta, and Stefanos Nikolaidis. 2021. Two-stage clustering of human preferences for action prediction in assembly tasks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3487–3494.

[47] John D Patterson, Baptiste Barbot, James Lloyd-Cox, and Roger E Beaty. 2024. AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods* 56, 4 (2024), 3619–3636.

[48] Mykola Pechenizkiy, Nikola Trcka, Ekaterina Vasilyeva, Wil MP van der Aalst, and PME De Bra. 2009. Process mining online assessment data. In *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining: proceedings [EDM'09], Cordoba, Spain. July 1-3, 2009*. International Working Group on Educational Data Mining, 279–288.

[49] Oksana Pylypchuk, Andrii Polubok, Olga Krivenko, Olena Safronova, Danylo Kosenko, and Nataliia Avdieieva. 2021. Developing an Approach to Colour Assessment of Works of Art on Aim to Creating a Comfortable and Harmonious Interior. In *2021 International Conference on Social Sciences and Big Data Application (ICSSBDA 2021)*. Atlantis Press, 181–187.

[50] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial images: using image prompts to improve subject representation in multimodal ai generated art. In *Proceedings of the 14th Conference on Creativity and Cognition*. 15–28.

[51] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[52] J Rowe, L Hughes, D Eckstein, and AM Owen. 2008. Rule-selection and action-selection have a shared neuroanatomical basis in the human prefrontal and parietal cortex. *Cerebral cortex* 18, 10 (2008), 2275–2285.

[53] Michelle J Searle and Lyn M Shulha. 2016. Capturing the imagination: Arts-informed inquiry as a method in program evaluation. *Canadian Journal of Program Evaluation* 31, 1 (2016), 34–60.

[54] Woosuk Seo, Joonyoung Jun, Minki Chun, Hyeonhak Jeong, Sungmin Na, Woohyun Cho, Saeri Kim, and Hyunggu Jung. 2022. Toward an AI-assisted Assessment Tool to Support Online Art Therapy Practices: A Pilot Study.. In *ECSCW*.

[55] S Sfarra, C Ibarra-Castanedo, D Ambrosini, D Paoletti, A Bendada, and X Maldague. 2014. Discovering the defects in paintings using non-destructive testing (NDT) techniques and passing through measurements of deformation. *Journal of Nondestructive Evaluation* 33 (2014), 358–383.

[56] Ping Shi, Peter B Luh, and David L Kleinman. 1992. The Impact of Individual and Team Goals on Human Distributed Dynamic Decisionmaking: A Mathematical Model. In *1992 American Control Conference*. IEEE, 1976–1977.

[57] Heng-Jie Song, Zhi-Qi Shen, Chun-Yan Miao, Ah-Hwee Tan, and Guo-Peng Zhao. 2007. The multi-agent data collection in HLA-based simulation system. In *21st International Workshop on Principles of Advanced and Distributed Simulation (PADS'07)*. IEEE, 61–69.

[58] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314* (2023).

[59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[60] Nikola Trcka and Mykola Pechenizkiy. 2009. From local patterns to global models: Towards domain driven educational process mining. In *2009 Ninth international conference on intelligent systems design and applications*. IEEE, 1114–1119.

[61] Mehrnoosh Vahdat, Luca Oneto, Davide Anguita, Mathias Funk, and Matthias Rauterberg. 2015. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15-18, 2015, Proceedings 10*. Springer, 352–366.

[62] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).

[63] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* (2023).

[64] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel Understanding with Visual Instruction Tuning. arXiv:2312.10032 [cs.CV] https://arxiv.org/abs/2312.10032

[65] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).

[66] Jun Zhang. 2020. *Two Systems of Chinese Life Aesthetics.* People's Publishing House, Beijing. 010–011 pages.

[67] Jiajing Zhang, Yongwei Miao, and Jinhui Yu. 2021. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access* 9 (2021), 77164–77187.

[68] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems* 36 (2023), 5484–5505.

[69] Liang Zhao, Eslam Hussam, Jin-Taek Seong, Assem Elshenawy, Mustafa Kamal, and Etaf Alshawarbeh. 2024. Revolutionizing art education: Integrating AI and multimedia for enhanced appreciation teaching. *Alexandria Engineering Journal* 93 (2024), 33–43.

[70] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations. arXiv:2207.00221 [cs.CV] https://arxiv.org/abs/2207.00221

A SYSTEM SETTINGS

A.1 Server

Our system is built using Python and Flask and is deployed locally. The server specifications are as follows:

- **Processor**: Intel Core i7-11800H
- **Memory**: 16GB DDR4 RAM
- **Graphics**: NVIDIA GeForce RTX 3060
- **Storage**: 512GB NVMe SSD
- **Operating System**: Windows 10

A.2 Decoding Parameters

We used the following decoding parameters for configuring the inference process of the GPT-4o model:

- **Engine**: GPT-4o
- **Response Length (Word Piece)**:
  – Entity Agent: 100
  – Review and Suggestion Agents: 500
- **Temperature**: 0 (This parameter controls the randomness of the generated text, with 0 indicating deterministic generation.)
- **Top P**: 1 (This value controls the diversity of candidate words during generation.)

A.3 Prompt Configuration

When generating outputs for the Entity Agent and Review Agent, we utilized specially designed prompts to guide the model. Below are the details of the prompts for each agent:

A.3.1 Entity Agent Prompt. The following prompt was used for the Entity Agent to identify and list objects or features in an image:
> **Prompt:**
>
> Identify and list the objects or features present in the image using descriptive labels. Use simple, clear terms like 'Face', 'Black hair', 'Thick lips', 'Big ears', etc. Ensure that each label is descriptive and that labels

are separated by the symbol (';'). For example: Face; Black hair; Thick lips; Big ears;. Also, identify the art style of the image with a label starting with '## Style:'.

A.3.2 Review Agent Prompts. The Review Agent generates evaluations based on nine distinct assessment dimensions, with each dimension corresponding to a specific prompt. Below a example of prompts for selected dimensions:

Table 1: Assessment Criteria for Realistic Artwork

Criterion: Realistic. This criterion assesses the accuracy of proportions, textures, lighting, and perspective to create a lifelike depiction.	
5:	The artwork exhibits exceptional detail and precision in depicting realistic features. Textures and lighting are used masterfully to mimic real-life appearances with accurate proportions and perspective. The representation is strikingly lifelike, demonstrating advanced skills in realism.
4:	The artwork presents a high level of detail and accuracy in the portrayal of subjects. Proportions and textures are very well executed, and the lighting enhances the realism. Although highly realistic, minor discrepancies in perspective or detail might be noticeable.
3:	The artwork represents subjects with a moderate level of realism. Basic proportions are correct, and some textures and lighting effects are used to enhance realism. However, the depiction may lack depth or detail in certain areas.
2:	The artwork attempts realism but struggles with accurate proportions and detailed textures. Lighting and perspective may be inconsistently applied, resulting in a less convincing depiction.
1:	The artwork shows minimal attention to realistic details. Proportions, textures, and lighting are poorly executed, making the depiction far from lifelike.

A.3.3 Suggestion Agent Prompt. The following table outlines the logic used in the Python function that dynamically generates prompts for the Suggestion Agent. This function is designed to provide improvement suggestions for each dimension based on the user-provided data: This function is used to dynamically generate prompts for the Suggestion Agent, providing suggestions for improvements in various dimensions.

For a full list of prompts and detailed information, please refer to the following GitHub repository: https://github.com/ArtMentor/ArtMentorApp/blob/main/ArtMentor_app.py

B INTERVIEW QUESTIONNAIRE

Greeting

Dear Participant,

Thank you for taking part in this interview. Your valuable insights will help us further optimize the ArtMentor system, enhancing its role in the field of art education. We are eager to learn about

Table 2: Prompt Generation Logic for Suggestion Agent

Step	Logic Description
Entity Extraction	Extracts entities from the 'labels_data["original"]' list and updates them by removing entities listed in 'labels_data["removed"]'. Newly added entities from 'labels_data["added"]' are inserted into the updated list. The result is concatenated into a string of entity labels separated by semicolons.
Score Review	Generates a prompt based on the score and review submitted by the user. The current score is extracted from 'score_Review_data["scores"]["current"]', and the current review is extracted from 'score_Review_data["Reviews"]["current"]'. These are combined into a single prompt string.
Suggestion Inclusion	Incorporates the current suggestion provided by the user into the prompt. The suggestion is extracted from 'suggestion_data["suggestions"]["current"]'.
Final Prompt Construction	Combines all previous components into the final prompt. The dimension name ('dimension') is included, and a message is added to inform the Suggestion Agent to consider user feedback and focus on improving the specified dimension. The final prompt instructs the model to output a dictionary format for ease of processing.

Main Interview Questions

(1) What are your thoughts on the accuracy of the system's entity recognition in artworks? Do you believe the system met your expectations in this regard?
(2) Regarding the overall usability of the system, how would you rate your experience with the human-computer interaction? If there are any areas for improvement, please provide detailed suggestions.
(3) Do you believe that the multimodal large language model effectively assisted you in scoring, commenting, and generating suggestions for the evaluation of artworks? Please briefly explain your reasoning.
(4) During your use of the system, which functional dimensions did you find to be particularly well-executed, and which ones do you think require improvement?
(5) Do you consider the suggestions provided by the system to be appropriate for the artistic capabilities of elementary students? Do these suggestions have the potential to inspire creativity and enhance the students' artistic skills?

Closing Remarks

Once again, thank you for your participation and support! Your feedback will directly contribute to the ongoing improvement of the ArtMentor system, helping us better meet the needs of our users. If you have any additional suggestions or further comments, please do not hesitate to contact us. We look forward to collaborating with you again in the future. Wishing you continued success in your work!

your experiences and suggestions while using the system. Below are some questions we would like to understand from your perspective. We appreciate you taking the time to provide thoughtful responses.

Basic Information

To better understand your background and how it might influence your interaction with the ArtMentor system, we would appreciate if you could provide the following basic information:

(1) What is your current role in the field of art education? (e.g., Teacher, Art Mentor, School Administrator, etc.)
(2) How many years of experience do you have in art education?
(3) Have you used other educational technology tools before? If yes, please briefly describe your experience with them.
(4) How often do you integrate technology into your teaching or mentoring activities? (e.g., Daily, Weekly, Occasionally)
(5) What are your main goals when using the ArtMentor system? (e.g., Enhancing student creativity, Providing personalized feedback, etc.)