# Dimension-Aware Active Annotation for Aesthetic Perception via Multi-Agent Human–AI Collaboration

**Ye Zhang[1], Jinlong He[1], Dongjie Wang[2*], Yupeng Zhou[1], Minghao Yin[1,3*]**

[1] College of Information Science and Technology, Northeast Normal University
[2] Electrical Engineering and Computer Science Department, University of Kansas
[3] Key Laboratory for Applied Statistics of Ministry of Education, Northeast Normal University
zhangy923@nenu.edu.cn, hejl104@nenu.edu.cn, wangdongjie@ku.edu, zhouyp605@nenu.edu.cn, ymh@nenu.edu.cn

## Abstract

To cultivate students' aesthetic development, teachers must objectively interpret and evaluate the artistic qualities and emotional resonance within their paintings—a process known as aesthetic perception. This evaluation process is labor-intensive and susceptible to biases due to variations among individual teachers. Advances in artificial intelligence (AI) motivate the use of AI-driven models to automate and enhance this aesthetic perception task. However, building effective AI-driven aesthetic perception models requires extensive datasets, which are typically labor-intensive and costly to gather. To address this, we propose a novel framework that selectively identifies the most challenging dimensions of aesthetic perception for expert annotation, using AI-generated pseudo-annotations to reduce cost and improve model performance. Our framework integrates a multi-agent active learning strategy to systematically annotate scores across multiple dimensions of aesthetic perception. Initially, we train an aesthetic perception model using a small, manually annotated dataset, establishing primary annotation capabilities. Then, this trained model generates pseudo-annotations for unlabeled data across various aesthetic dimensions (e.g., humor, happiness). To ensure annotation quality and relevance, a multi-agent system evaluates these pseudo-annotations, identifying dimensions requiring expert human input based on metrics such as model estimation confidence. Human experts provide targeted annotations selectively, refining the dataset and guiding an iterative improvement cycle. Through repeated refinement, the model progressively enhances both its predictive accuracy and its automated annotation proficiency. Our optimization approach dynamically balances accuracy, annotation relevance, and human effort. Extensive experiments conducted on two real-world datasets demonstrate the effectiveness of our framework.

## Introduction

To foster students' aesthetic and cognitive development, teachers must objectively interpret and evaluate the artistic qualities and emotional resonance present in their artwork, enabling effective and constructive feedback. This critical process, termed aesthetic perception, demands significant time and effort from educators. However, the subjectivity inherent in aesthetic evaluations often introduces biases stemming from variations in individual teachers' experiences. Traditional methods often struggle to systematically address these complexities, making it challenging to provide standardized and high-quality assessments consistently.

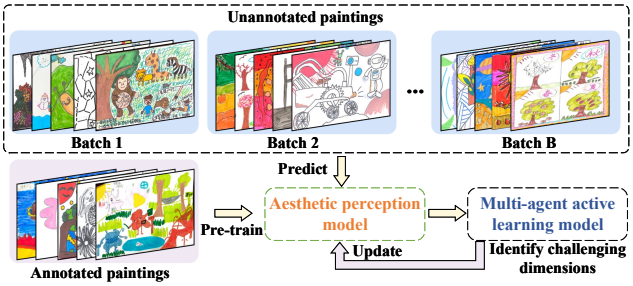



Figure 1: Human-AI collaboration decides which dimensions of student paintings require human annotation to enhance the performance of the aesthetic perception model.

In recent years, artificial intelligence (AI) has seen rapid development, leading to its widespread integration in education as a means to enhance the understanding and analysis of visual expression (Huang et al. 2025; Zhang et al. 2025a). Contemporary AI models are particularly adept at extracting intricate visual and emotional features from paintings, enabling deeper and more consistent interpretations compared to traditional human assessments. For instance, deep learning techniques have demonstrated significant success in tasks such as aesthetic quality assessment and emotional recognition in artwork (Lu et al. 2015; Kao, He, and Huang 2017). By leveraging large-scale datasets, these AI-driven approaches facilitate personalized and scalable solutions, presenting novel opportunities to address the inherent limitations of conventional educational evaluation practices.

To construct such an evaluation model, extensive and high-quality labeled data are essential. But, in this domain, there is a lack of sufficient labeled data, and the annotation process is typically time-consuming and costly. Prior studies have sought to mitigate this challenge using techniques such as semi-supervised learning and transfer learning (Wang and Shen 2017; Zhang, Miao, and Yu 2021; Zhang et al. 2024). Although these methods partially alleviate the demand for fully annotated datasets, they still rely heavily on substantial labeled data and often struggle to deliver precise annotations across multiple dimensions of aesthetic perception.

*Corresponding author

Therefore, our aim is to develop an accurate aesthetic perception model while reducing the human effort required to construct a large-scale labeled dataset. To this end, we propose a **H**uman-**AI** collaborative **P**ainting **A**nnotation (**HAPA**) framework, which automatically evaluates paintings across multiple aesthetic dimensions and identifies the most challenging dimensions for targeted human annotation. By focusing expert effort where it is most needed, our framework enhances model performance while significantly reducing annotation costs and human workload. Figure 1 illustrates the high-level pipeline of HAPA. More specifically, we begin by using a small amount of labeled data to initialize the aesthetic perception model. This initial model is then employed to generate pseudo-annotations for unlabeled paintings in a batch-wise manner. For each incoming batch, the model estimates scores across various aesthetic dimensions (e.g., humor, happiness). To determine which dimensions require expert intervention, we introduce a reinforcement learning (RL) agent for each dimension. These agents collectively form a multi-agent evaluation system that assesses the model's predictions and identifies the most challenging dimensions—those with poor predictive performance—for human annotation. The overall objective of this process is to optimize a trade-off: maximizing the perception model's accuracy and confidence while minimizing the cost of human annotation. Once human experts annotate the selected dimensions in a batch, the newly labeled data are added to the training set to incrementally retrain and improve the aesthetic perception model. As the model evolves, its ability to accurately and reliably annotate each aesthetic dimension improves. Simultaneously, the accumulation of high-quality labeled data enhances the stability and precision of the learning process. Through this iterative and targeted strategy, the proposed HAPA framework facilitates the construction of a high-performance aesthetic perception model with minimal reliance on human annotation effort.

To summarize, our main contributions are as follows:

- We propose a novel human–AI collaborative framework for aesthetic perception that strategically enhances model accuracy while significantly reducing annotation costs. Our approach provides a general paradigm for human-in-the-loop learning in multi-dimensional perception tasks.
- We develop a multi-agent active learning framework, in which each agent is responsible for a specific aesthetic dimension and learns to determine whether human annotation is needed by balancing model uncertainty and annotation cost. This design enables fine-grained control and scalability across diverse perception dimensions.
- We conduct extensive experiments to validate the effectiveness, robustness, and interpretability of our framework. Results show substantial cost reduction and consistent performance gains, highlighting the HAPA's practicality and broader applicability.

## Problem Formulation and Definitions

Our objective is to locate the most challenging aesthetic dimensions whose human annotation yields the greatest enhancement in the performance of the aesthetic perception model. Formally, given a set of aesthetic perception dimensions $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ and an unlabeled student painting $\mathcal{P}$, our objective is to identify a subset of challenging dimensions $\mathcal{D}^* \subseteq \mathcal{D}$ for human annotation, such that the resulting update to the aesthetic perception model $M$ yields the greatest overall performance improvement. In this setting, dimensions in $\mathcal{D}^*$ are labeled by human experts, while the remaining dimensions $\mathcal{D} \setminus \mathcal{D}^*$ are automatically annotated by the current model $M$. This selective annotation strategy aims to maximize the expected gain in model perception accuracy while minimizing human annotation effort. We define the optimal subset $\mathcal{D}^*$ as follows:

$$\mathcal{D}^* = \arg\max_{\mathcal{D}' \subseteq \mathcal{D}} V(\mathcal{D}', \mathcal{P}), \tag{1}$$

where $V(\mathcal{D}', \mathcal{P})$ denotes the optimal performance of the perception model after human experts annotate $\mathcal{D}'$ and the model auto-annotates $\mathcal{D} \setminus \mathcal{D}'$ for painting $\mathcal{P}$.

To achieve this goal, we propose a multi-agent reinforcement learning-based active learning framework that identifies and selects the most challenging perception dimensions requiring human annotation. Multiple agents collectively select $\mathcal{D}^*$ based on improving perception accuracy and reducing the cost of annotation. **Agent.** Given $N$ aesthetic perception dimensions, $N$ agents are assigned, each agent responsible for making the decision on its corresponding dimension. **Action.** For each agent, the action space is defined as $a \in \{0, 1\}$, where 1 denotes the dimension is manually annotated, and 0 denotes it is annotated by the aesthetic perception model. **State.** The state $s$ of each agent encodes the current dimension subset selection, providing contextual information for aesthetic perception evaluation. **Reward.** The reward $r$ reflects the effectiveness of the selected dimension subset, combining (1) the performance improvement of the aesthetic perception model from human annotations (posi-

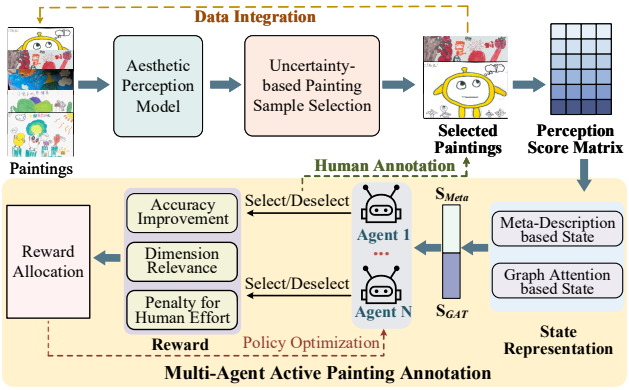



Figure 2: An overview of HAPA. We initially employ an aesthetic perception model to evaluate student paintings across multiple perception dimensions. Then, each dimension is managed by a dedicated agent responsible for determining whether it requires human annotation. Descriptive statistics and attention-based graph networks encode the current annotation state, enabling agents to learn optimal dimension-selection policies guided by a reward allocation strategy. The iterative process is designed to minimize annotation costs while maximizing perception accuracy.

tive reward) and (2) the total annotation cost (negative reward). This reward is then distributed among agents based on their individual contributions to the selection.

# Methodology

## Framework Overview

Figure 2 illustrates our proposed Human-AI collaborative Painting Annotation (HAPA) framework. It employs a multi-agent active learning (MAAL) strategy to optimize annotation efficiency and enhance aesthetic perception performance. Initially, HAPA trains an aesthetic perception model with limited annotated data, enabling automated annotations across multiple aesthetic dimensions. During the active learning phase, this trained model generates pseudo-annotations for unlabeled data. A multi-agent reinforcement learning (MARL) component strategically selects specific aesthetic dimensions requiring additional human annotation based on perception difficulty and dimensionality relevance. Human experts annotate these selected dimensions, updating the aesthetic model accordingly. This iterative MAAL process significantly reduces annotation effort while progressively improving the perception model accuracy.

## Aesthetic Perception Model

As shown in Figure 3, aesthetic perception model serves two purposes: (1) quantitatively assessing students' aesthetic capability across multiple dimensions; (2) generating pseudo-labels for unlabeled student paintings. The model has two main components: a global aesthetic feature encoder that integrates CNNs and Transformers to capture color and emotional features from student paintings; a set of dimension-aligned evaluation heads, each tailored to evaluate a specific aesthetic dimension based on the shared representation.

**Global Aesthetic Feature Encoder.** To extract both color and emotional cues from student paintings, we design a hybrid encoder combining Convolutional Neural Networks (CNNs) and a Transformer. The CNN module comprises two convolutional layers followed by two pooling layers. The first block captures low-level visual features, such as edges and textures, while the second block extracts higher-level attributes related to color distribution and compositional structure. The resulting feature maps are flattened and fed into a Transformer encoder, which models global dependencies and captures emotional semantics. The output is a unified representation vector $\boldsymbol{h}_{\text{global}}$ that encodes both the chromatic and affective characteristics of the input painting.

**Dimension-Aligned Evaluation Heads.** To evaluate various aspects of aesthetic perception, we employ a set of dimension-aligned evaluation heads, each dedicated to a specific aesthetic dimension (e.g., humor, curiosity, pain). While these heads share a consistent architectural design—comprising two fully connected (FC) layers—they are independently parameterized and trained to specialize in their respective perceptual targets. Each evaluation head takes the shared representation vector $\boldsymbol{h}_{\text{global}}$ as input and produces a scalar prediction corresponding to its assigned

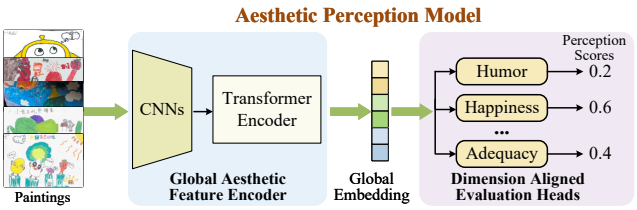



Figure 3: The aesthetic perception model consists of a global encoder for visual-emotional features and dimension-aligned heads for multi-dimensional perception scoring.

aesthetic dimension. Formally, the output of the $i$-th evaluation head is given by: $\hat{y}_i = \text{Head}_i(\boldsymbol{h}_{\text{global}})$, where $\text{Head}_i$ denotes the $i$-th evaluation head, and $\hat{y}_i$ represents the predicted score for the $i$-th aesthetic dimension. By incorporating these heads, the model supports multi-dimensional aesthetic assessment in a unified and scalable manner.

**Optimization.** To effectively train the aesthetic perception model, we propose a joint optimization strategy that updates both the global aesthetic feature encoder and the dimension-aligned evaluation heads. Our approach targets two key challenges: (1) learning from under-represented extreme perception scores, and (2) ensuring each evaluation head specializes in its corresponding aesthetic dimension.

To tackle the first challenge, we introduce a difficulty-aware loss that focuses training on hard-to-predict samples. Specifically, we adopt a focal-style regression loss that increases the contribution of samples with large prediction errors. For the $i$-th aesthetic dimension, this loss is defined as:

$$\mathcal{L}_{\text{DA}}^{(i)} = (1 - \exp(-|y_i - \hat{y}_i|))^{\eta} \cdot (y_i - \hat{y}_i)^2, \qquad (2)$$

where $y_i$ is the ground-truth score, $\hat{y}_i$ is the predicted score, and $\eta > 0$ controls the focusing strength. When the prediction error $|y_i - \hat{y}_i|$ is small, the loss is suppressed; when the error is large, the loss is amplified, thereby guiding the model to focus on difficult samples, especially those from under-represented perception score regions.

To address the second challenge, we introduce a dimension-specific regression loss to ensure accurate and independent modeling of each aesthetic dimension, defined as the Mean Absolute Error (MAE):

$$\mathcal{L}_{\text{Reg}}^{(i)} = |y_i - \hat{y}_i|. \qquad (3)$$

The loss for the $i$-th dimension-specific evaluation head combines the difficulty-aware loss and the MAE loss. During training, each evaluation head is optimized using its own loss to specialize in the corresponding aesthetic dimension. Gradients from all heads are aggregated and backpropagated to the shared encoder, enabling the model to learn both global embeddings and accurate dimension predictions.

## Multi-Agent Active Painting Annotation

Manual annotation of perceptual data is costly and labor-intensive, especially when the annotation space involves multiple aesthetic dimensions. To address this, we propose a **multi-agent active annotation framework** that strategically identifies (1) the most uncertain painting samples and (2) the most ambiguous perceptual dimensions within

each selected sample. This two-stage selection process is guided by uncertainty estimation and multi-agent reinforcement learning (MARL), aiming to minimize human annotation cost while maximizing perceptual modeling accuracy.

**Stage I: Uncertainty-Based Painting Sample Selection.** Let $x$ denote an unlabeled painting sample from a given batch. Using the previously described perception model, we obtain a trained model $f_\theta$ that produces multi-dimensional predictions $\hat{\boldsymbol{y}} = f_\theta(x) \in \mathbb{R}^N$ for the sample $x$, where $N$ is the number of perceptual dimensions. To identify samples for which the model exhibits the lowest confidence, we compute an uncertainty score $u(x)$ for each sample based on the prediction uncertainty across the predicted dimensions: $u(x) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{H}(\hat{\boldsymbol{y}}^j)$, where $\mathcal{H}(\cdot)$ denotes the uncertainty measure (e.g., empirical variance across ensemble predictions). The top-$K$ samples with the highest uncertainty score are selected to form the candidate batch $\mathcal{B}$ for dimension-level exploration.

**Stage II: Multi-Agent Dimension Selection.** In this stage, we introduce a multi-agent reinforcement learning (MARL) framework to strategically identify which perceptual dimensions should be annotated for each selected sample. The objective is to maximize model improvement while minimizing unnecessary human effort. To this end, the learning process is guided by three principles: enhancing prediction accuracy, promoting semantic relevance among selected dimensions, and penalizing excessive annotation.

For each sample $x \in \mathcal{B}$, the aesthetic perception model generates a multi-dimensional prediction vector in $\mathbb{R}^N$, where $N$ denotes the number of perceptual dimensions. Correspondingly, we create a set of $N$ agents (Hester et al. 2018), with each agent dedicated to a specific dimension. Each agent decides whether its assigned dimension should be selected for human annotation, based on a learned policy that integrates both statistical patterns and structural relationships among dimensions.

**State Representation.** To enable agents to make more informed decisions, we develop two methods to derive state representations: meta descriptive statistics, and GAT-based representation. These methods capture both individual dimension characteristics and their interrelationships, which is crucial for MARL agents to make informed decisions. *1) Meta Descriptive Statistics*. For current selected aesthetic dimension, we compute seven statistical descriptors: mean, standard deviation, minimum, maximum, and the first, second, and third quartiles, forming a $1 \times 7$ vector representing the overall statistical profile, denoted as $\mathcal{S}_{meta}$. This structured representation provides each agent with a compact and informative global view, facilitating an efficient assessment of the distribution patterns among aesthetic perception dimensions. *2) GAT-based representation*. Based on the hierarchical structure of aesthetic perception dimensions (Zhang et al. 2024), we construct a graph $G$ where each node represents a dimension and edges encode semantic relevance between dimensions. We then apply a Graph Attention Network (GAT) (Velickovic et al. 2017) to compute attention coefficients that quantify the influence of neighboring nodes, allowing each node's embedding to be updated through weighted aggregation. The GAT-based state representation $\mathcal{S}_{GAT}$ captures the structural interdependencies among current selected dimensions. Finally, we concatenate $\mathcal{S}_{meta}$ and $\mathcal{S}_{GAT}$ to form the state at the $t$-th iteration $s_i^t = [\mathcal{S}_{meta} \,||\, \mathcal{S}_{GAT}]$.

**Reward.** To promote efficient and balanced decision-making in the human-AI collaboration process, we design a reward mechanism that encourages agents to select dimensions that 1) improve prediction accuracy; 2) ensure semantic relevance; and 3) minimize unnecessary human intervention. The total reward $r^{t+1}$ at iteration $t+1$ is computed as a weighted sum of three components: accuracy improvement ($\mathcal{R}_{acc}$), dimension relevance ($\mathcal{R}_{rel}$), and a penalty for annotation cost ($\mathcal{R}_{pen}$). 1) *Accuracy Improvement* ($\mathcal{R}_{acc}$). This term quantifies the gain in model performance based on the reduction in Mean Squared Error (MSE) after incorporating human-provided annotations: $\mathcal{R}_{acc} = \mathcal{L}_t - \mathcal{L}_{t+1}$, where $\mathcal{L}_t$ and $\mathcal{L}_{t+1}$ denote the MSE values before and after annotation, respectively. A positive value indicates improved prediction accuracy. 2) *Dimension Relevance* ($\mathcal{R}_{rel}$). To ensure that the selected dimensions are meaningful and complementary, we compute their pairwise Pearson correlations based on GAT-derived embeddings: $\mathcal{R}_{rel} = \sum_{d_1,d_2 \in \text{DS}} \text{Pearson}(\mathbf{e}_{d_1}, \mathbf{e}_{d_2})$, where DS is the set of selected dimensions and $e_{d_i}$ is the representation of the $d_i$-th dimension. 3) *Penalty for Human Effort* ($\mathcal{R}_{pen}$). To discourage excessive reliance on manual annotation, we introduce a penalty proportional to the number of agents that request human annotation in the current iteration. The overall reward is computed as: $r^{t+1} = \alpha \cdot \mathcal{R}_{acc} + \beta \cdot \mathcal{R}_{rel} - \gamma \cdot \mathcal{R}_{pen}$, where $\alpha$, $\beta$, and $\gamma$ are hyperparameters that control the trade-offs among perception accuracy, dimensional relevance, and human annotation cost.

**Reward Allocation Strategy.** To promote fair collaboration and discourage redundant annotation requests, we design a reward allocation strategy that incentivizes deliberate and confident actions. Let $m_t = \sum_i a_i^t$ denote the number of agents that requested human annotation at time step $t$, where $a_i^t = 1$ indicates selection. If $m_t > 0$, the total reward $r^{t+1}$ is distributed equally among these participating agents; otherwise, all agents receive zero reward. This allocation mechanism discourages excessive or low-confidence selections by reducing the reward share when too many agents request annotation. It thereby reinforces a balance between annotation efficiency and model improvement.

**Policy Optimization.** To enable agents to learn effective dimension-selection strategies over time, we optimize their decision-making policies using Deep Q-Learning. The goal is to minimize the difference between the predicted Q-values and the target values derived from actual rewards and future estimates. The policy loss is defined as:

$$\mathcal{L}_{\text{policy}} = \frac{1}{N}\sum_{i=1}^{N} \left(Q(s_i^t, a_i^t \mid \theta_t) - y_i^t\right)^2 \tag{4}$$

where the target Q-value is computed as: $y_i^t = r_i^t + \Gamma \cdot \max_{a_i^{t+1}} Q(s_i^{t+1}, a_i^{t+1} \mid \theta_{t+1})$. Here, $\Gamma$ is the discount factor, and $\theta_t$ denotes the parameters of the Q-network at iteration $t$. By minimizing this loss via gradient descent, each

| Baselines | CCPS | | | | | | BCPS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | RMSE | MSE ↓ | MAE ↓ | RMSE ↓ | MSE | MAE | RMSE | MSE ↓ | MAE ↓ | RMSE ↓ |
| CNN | 0.2213 | 0.4117 | 0.4705 | 73.57% | 55.14% | 48.59% | 0.1705 | 0.3399 | 0.4130 | 59.71% | 42.57% | 36.51% |
| ResNet | 0.2292 | 0.4193 | 0.4787 | 74.48% | 55.95% | 49.47% | 0.1863 | 0.3532 | 0.4317 | 63.12% | 44.73% | 39.26% |
| AlexNet | 0.2967 | 0.4896 | 0.5447 | 80.28% | 62.28% | 55.59% | 0.2616 | 0.4227 | 0.5115 | 73.74% | 53.82% | 48.74% |
| VGGNet | 0.2902 | 0.4832 | 0.5387 | 79.84% | 61.78% | 55.10% | 0.3124 | 0.4734 | 0.5589 | 78.01% | 58.77% | 53.09% |
| SqueezeNet | 0.2695 | 0.4625 | 0.5192 | 78.29% | 60.06% | 53.41% | 0.2761 | 0.4371 | 0.5254 | 75.12% | 55.34% | 50.10% |
| DenseNet | 0.2742 | 0.4673 | 0.5236 | 78.67% | 60.48% | 53.80% | 0.2393 | 0.4003 | 0.4892 | 71.29% | 51.24% | 46.40% |
| VAN | 0.2327 | 0.4250 | 0.4824 | 74.86% | 56.54% | 49.85% | 0.1780 | 0.3409 | 0.4218 | 61.40% | 42.74% | 37.84% |
| MixMAE | 0.2359 | 0.4276 | 0.4856 | 75.20% | 56.81% | 50.19% | 0.1881 | 0.3578 | 0.4337 | 63.48% | 45.44% | 39.54% |
| Rand | 0.0705 | 0.2033 | 0.2656 | 17.02% | 9.15% | 8.92% | 0.0742 | 0.2053 | 0.2724 | 7.41% | 4.92% | 3.74% |
| Kmeans++ | 0.0720 | 0.2043 | 0.2683 | 18.75% | 9.59% | 9.84% | 0.0736 | 0.2049 | 0.2714 | 6.66% | 4.73% | 3.39% |
| Margin | 0.0698 | 0.1982 | 0.2642 | 16.19% | 6.81% | 8.44% | 0.0769 | 0.2040 | 0.2773 | 10.66% | 4.31% | 5.45% |
| Entropy | 0.0643 | 0.1926 | 0.2535 | 9.02% | 4.10% | 4.58% | 0.0765 | 0.2040 | 0.2766 | 10.20% | 4.31% | 5.21% |
| Coreset | 0.0627 | 0.1934 | 0.2503 | 6.70% | 4.50% | 3.36% | 0.0755 | 0.2144 | 0.2748 | 9.01% | 8.96% | 4.59% |
| BatchBALD | 0.0647 | 0.1914 | 0.2543 | 9.58% | 3.50% | 4.88% | 0.0762 | 0.2072 | 0.2761 | 9.84% | 5.79% | 5.03% |
| BADGE | 0.0617 | 0.1914 | 0.2483 | 5.19% | 3.50% | 2.58% | 0.0721 | 0.1956 | 0.2686 | 4.72% | 0.20% | 2.38% |
| Bait | 0.0629 | 0.1932 | 0.2509 | 7.00% | 4.40% | 3.59% | 0.0773 | 0.2073 | 0.2781 | 11.13% | 5.84% | 5.72% |
| VeSSAL | 0.0749 | 0.2071 | 0.2737 | 21.90% | 10.82% | 11.62% | 0.0712 | 0.1953 | 0.2667 | 3.51% | 0.05% | 1.69% |
| HAPA | **0.0585** | **0.1847** | **0.2419** | — | — | — | **0.0687** | **0.1952** | **0.2622** | — | — | — |

Table 1: MSE, MAE and RMSE comparison of all baselines on CCPS and BCPS. ↓ indicate performance improvements.

agent incrementally improves its policy, enabling it to identify the most valuable dimensions for annotation in a cost-effective and performance-aware manner.

## Experimental Setup

**Dataset.** We evaluate the effectiveness of HAPA on a dataset of 5138 classroom paintings collected from two elementary schools in China: 2,889 from Changchun Primary School (CCPS) and 2,239 from Baicheng Primary School (BCPS). Each painting was independently evaluated by 16 art education experts. The scores were averaged and normalized to a [0, 1] range across multiple aesthetic perception dimensions, serving as ground-truth labels. For the active learning setup, we randomly split the dataset into 80% for training and 20% for testing. The test set is held out for performance evaluation. Within the training set, a small subset is initially labeled to train the perception model, while the remaining samples serve as the unlabeled pool. In each iteration, the model actively selects uncertain samples for annotation, and the labeled set is gradually expanded.

**Evaluation Metrics.** To evaluate the efficacy of HAPA, we consider two key aspects: perception accuracy and annotation cost. For perception accuracy, we report Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). where lower values indicate better performance. To assess annotation cost, we track the total number of human-annotated dimensions required during the active learning process.

**Baseline Algorithms.** To comprehensively evaluate the performance of our framework, we compare against two categories of baselines, each targeting a different aspect of the task: (1) the effectiveness of the aesthetic perception model, and (2) the efficacy of the proposed multi-agent active learning strategy. **Aesthetic Perception Model Baselines.** To assess the visual representation capability of our model, we compare it with eight representative image recognition methods, which are employed solely to extract visual features for aesthetic perception modeling. The baselines include: **CNN** (Chauhan, Ghanshala, and Joshi 2018), a standard convolutional architecture; **ResNet** (Targ, Almeida, and Lyman 2016), a deep residual network with skip connections; **AlexNet** (Yuan and Zhang 2016), an early deep CNN utilizing ReLU and dropout; **VGGNet** (Simonyan and Zisserman 2014), which stacks small convolutional filters for spatial consistency; **SqueezeNet** (Iandola et al. 2016), a lightweight model with fewer parameters; **DenseNet** (Huang et al. 2017), which densely connects layers to improve gradient flow; **VAN** (Guo et al. 2023), which integrates attention mechanisms for enhanced feature localization; and **MixMAE** (Liu et al. 2023), a self-supervised masked image modeling method for learning interpretable features. **Active Learning Strategy Baselines.** To validate the effectiveness of our Multi-Agent Active Learning (MAAL) module, we compare it with nine widely used active learning strategies integrated into the same backbone framework. These include: **Rand**, which selects dimensions randomly and serves as a naive baseline; **Kmeans++** (Arthur and Vassilvitskii 2006), which promotes sample diversity via clustering-based initialization; **Margin** (Roth and Small 2006), which selects samples with the smallest prediction margins; **Entropy** (Wang and Shang 2014), which prioritizes high-uncertainty samples; **Core-set** (Sener and Savarese 2017), which identifies representative samples for improved generalization; **BatchBALD** (Kirsch, Van Amersfoort, and Gal 2019), which selects informative batches by maximizing mutual information; **BADGE** (Ash et al. 2019), which balances uncertainty and gradient-based diversity; **BAIT** (Ash et al. 2021), which minimizes MLE error bounds using Fisher information; and **VeSSAL** (Saran et al. 2023), which performs efficient and adaptive streaming batch selection with automatic uncertainty-diversity balancing.

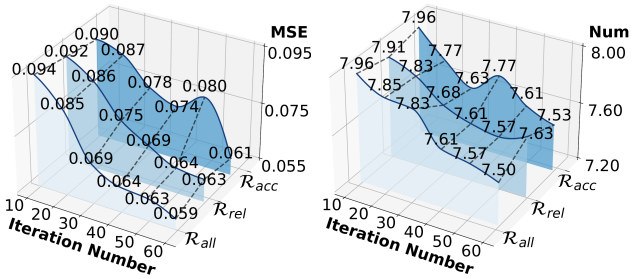
## Experimental Results

**Overall Performance.** This experiment is conducted to evaluate the overall effectiveness of HAPA in enhancing aesthetic perception performance while reducing annotation cost through multi-agent active learning. We compare

HAPA with conventional image recognition models and state-of-the-art active learning strategies on the CCPS and BCPS datasets using MSE, MAE and RMSE as evaluation metrics. As shown in Table 1, HAPA consistently outperforms all baselines in prediction accuracy. Compared to traditional models, it significantly reduces prediction error, and compared to advanced active learning methods, it achieves superior performance under the same annotation budget. These results show two advantages of our approach: (1) the multi-agent structure enables precise identification of hard-to-predict aesthetic dimensions, ensuring annotation is targeted and efficient; and (2) the framework delivers higher accuracy than competing methods given the same level of human annotation effort. Overall, HAPA offers a practical and effective solution for AI-assisted art education.

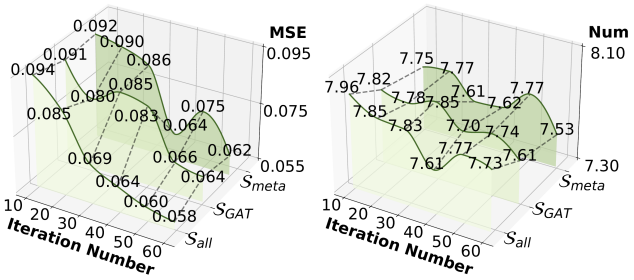
**Reward Function Analysis.** This experiment is conducted on CCPS dataset to evaluate the impact of different reward configurations of HAPA. In our reward design, the penalty term $\mathcal{R}_{pen}$ plays a critical role in controlling the number of human annotations, thereby ensuring that the annotation cost remains manageable. As such, $\mathcal{R}_{pen}$ is included in all reward configurations evaluated in this experiment. We compare three configurations: (i) $\mathcal{R}_{acc}$, which combines the accuracy reward with $\mathcal{R}_{pen}$; (ii) $\mathcal{R}_{rel}$, which combines the relevance reward with $\mathcal{R}_{pen}$; and (iii) $\mathcal{R}_{all}$, which integrates accuracy, relevance, and penalty terms. As shown in Figure 4, $\mathcal{R}_{all}$ achieves the best performance across all evaluation metrics. The superior performance of $\mathcal{R}_{all}$ can be attributed to its ability to jointly optimize prediction accuracy and perceptual relevance while maintaining low annotation cost. In contrast, $\mathcal{R}_{acc}$ and $\mathcal{R}_{rel}$ yield competitive but suboptimal results, as each focuses on only a single objective. Notably, because all configurations include $\mathcal{R}_{pen}$, the number of human annotations remains consistent, ensuring a fair comparison across settings. These results highlight the importance of holistic reward design in human-AI collaborative tasks. They reflect that incorporating multiple dimensions of model performance, rather than optimizing a single objective, leads to more robust and effective outcomes under constrained annotation budgets.

**State Representation Analysis.** This experiment is conducted on CCPS dataset to compare the effectiveness of different state representation strategies in HAPA. We evaluate three variants: (i) $\mathcal{S}_{meta}$, which encodes meta-level descriptive statistics; (ii) $\mathcal{S}_{GAT}$, which captures relational structure via a graph attention network; and (iii) $\mathcal{S}_{all}$, which combines both representations. Figure 5 shows that $\mathcal{S}_{all}$ consistently achieves the best performance across all evaluation metrics. Initially, $\mathcal{S}_{GAT}$ performs better than $\mathcal{S}_{meta}$ due to its ability to capture complex inter-dimensional relationships. However, after approximately 30 training iterations, $\mathcal{S}_{meta}$ begins to outperform $\mathcal{S}_{GAT}$, reflecting its growing effectiveness in modeling global statistical trends. By integrating both descriptive and relational information, $\mathcal{S}_{all}$ leverages their complementary strengths and further improves the model's predictive accuracy and stability. In addition, the number of human annotations stabilizes after around 10 iterations under the control of the reward mechanism, ensuring consistent annotation cost across settings. These findings highlight the importance of hybrid state representations. While $\mathcal{S}_{GAT}$ is advantageous in the early stages due to its fine-grained relational modeling, $\mathcal{S}_{meta}$ becomes increasingly robust over time. Their combination in $\mathcal{S}_{all}$ offers a more comprehensive and adaptive representation, making it well-suited for dynamic decision-making in human-AI collaborative environments.




(a) MSE (b) Human annotation number

Figure 4: Comparison of different HAPA reward.




(a) MSE (b) Human annotation number

Figure 5: Comparison of different HAPA state.

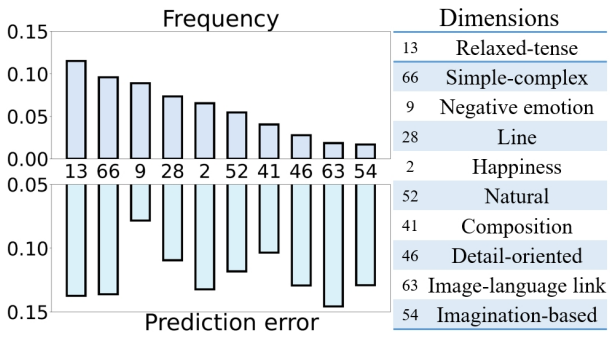



Figure 6: Top 10 dimensions by annotation frequency (above) and prediction error (below).

**Annotation Dimension Analysis.** This analysis is conducted to investigate which aesthetic perception dimensions most frequently require human annotation and to examine their corresponding prediction errors. As shown in Figure 6, the ten most frequently annotated dimensions include *Relaxed–tense*, *Simple–complex*, *Negative emotion*, *Line*, and *Happiness*, among others. Dimensions such as *Relaxed–tense* and *Simple–complex* appear with high frequency, indicating that stylistic attributes involving subtle interpretation and high subjectivity are particularly challenging for the model to predict. These dimensions rely

heavily on context, subjective judgment, and subtle artistic cues that are difficult to capture through visual features alone. On the other hand, emotional dimensions like *Negative emotion* and *Happiness* exhibit relatively high prediction errors, suggesting the model's limited ability to infer affective intent and cultural meaning embedded in students' artworks. The gap between predicted and actual scores highlights the inherent difficulty of modeling human emotion and experience using computational approaches. These findings indicate that while AI-based aesthetic perception models effectively learn structured visual patterns, they struggle to generalize to abstract, subjective, or emotionally rich dimensions. Human expertise remains indispensable for accurate, context-aware annotation, especially in tasks requiring deeper emotional or cultural insight.

## Related Works

**AI Empowerment Aesthetic Education** benefits greatly from the integration of AI technologies, which enhance the development of students' perceptive and creative abilities. Recent research emphasizes the importance of accurately identifying emotional and visual elements within artworks, facilitating deeper artistic engagement and interpretation (Bhandari, Chang, and Neben 2019; Whiting 2023). By guiding students to understand subtle expressions and underlying artistic contexts, AI-supported methods effectively foster richer aesthetic appreciation. Annotations of artwork have also been valuable in guiding students' artistic creation, helping them articulate originality and refine technical skills (Mao, Hong, and Nguyen 2023; Jin et al. 2022). Moreover, cognitive studies, particularly involving children, demonstrate that annotations provide meaningful insights into non-verbal emotional and cognitive development (Juanzi et al. 2019; Qiming and Renganathan 2024; Zhang et al. 2025b). Collectively, these perspectives highlight the significant potential of human-AI collaboration in supporting students' comprehensive aesthetic growth, ensuring emotional depth, cultural context, and creativity are appropriately recognized and nurtured within educational environments (Agboola and Yassin 2025; Zhang et al. 2024).

**Multi-Agent Reinforcement Learning** (MARL) focuses on optimizing interactions and coordination among multiple agents in a shared environment to maximize cumulative rewards (Chen et al. 2024). A dominant framework is centralized training with decentralized execution (CTDE), where agents exploit global information during training but make independent decisions at execution, balancing cooperation and autonomy (Agarwal et al. 2021). Algorithms such as Deep Q-Networks (DQN) and actor–critic methods are widely adopted for their effectiveness in handling shared rewards and ensuring stable convergence in cooperative multi-agent settings (Xiao et al. 2023; Wang et al. 2021). Recent advances integrate representation learning to extract compact, task-relevant features from high-dimensional inputs, improving performance in complex environments (Lin et al. 2018). Despite these advances, MARL suffers from slow convergence, especially in large action spaces. Reward shaping and hierarchical reinforcement learning have been explored to accelerate training while maintaining policy quality (Liu et al. 2019, 2025). Additionally, attention mechanisms and inter-agent communication protocols enhance coordination under dynamic or partially observable conditions (Yang et al. 2018). Improving the efficiency and scalability of MARL remains a critical research challenge.

**Active learning** has gained significant attention for its ability to reduce annotation costs by selecting the most informative samples for labeling. Recent research focuses on optimizing selection strategies to improve efficiency and adapt to domain-specific challenges. Batch selection methods that maximize mutual information and gradient diversity have shown notable advantages in identifying valuable samples (Rubashevskii, Kotova, and Panov 2023). Diversity-based approaches, such as probabilistic coverage and Bayesian estimation, further enhance performance in low-budget settings (Wang, Chen, and Du 2023; Melo et al. 2024). Applications in areas like biomedical image segmentation confirm that AL can significantly lower annotation efforts while maintaining high model accuracy (Gaillochet, Desrosiers, and Lombaert 2023). The integration of AL with semi-supervised and self-supervised techniques has gained traction, effectively combining labeled and unlabeled data to reduce costs without compromising performance (Gao et al. 2020). Addressing distribution shifts in biased datasets remains an ongoing challenge, with recent solutions focusing on improving model robustness through targeted sampling and reweighting strategies (Adachi et al. 2024).

## Conclusion

In this paper, we present a human-AI teaming framework for optimizing aesthetic perception annotation through the integration of automated modeling and targeted human intervention. The framework comprises two key components: an aesthetic perception model and a multi-agent active annotation system. The aesthetic perception model combines a global aesthetic encoder with dimension-specific evaluation heads, enabling the model to learn both general and fine-grained perceptual features. Trained initially on a small labeled dataset, the model generates pseudo-annotations for unlabeled data across multiple aesthetic dimensions. To improve annotation efficiency, we introduce a multi-agent active learning mechanism in which each agent is responsible for determining whether its corresponding dimension requires human annotation. The decision process is guided by a reward function that jointly optimizes model performance, annotation cost, and dimension relevance. Annotated data provided by human experts are incrementally incorporated into the model, allowing continuous refinement of perception accuracy. This collaborative learning cycle ensures that the system evolves through the complementary strengths of humans and AI. Experimental results demonstrate that our framework consistently outperforms strong baselines, achieving higher predictive performance with reduced human annotation effort. In future work, we aim to explore transfer learning to adapt the aesthetic perception model across diverse student groups and cultural contexts. This would enhance the model's generalizability while reducing annotation effort, ultimately supporting more inclusive and scalable applications in AI-assisted art education.

## Acknowledgments

This work was supported in part by the Jilin province science and technology department project under Grant 20240602005RC, NSFC under Grant No.(62206045, 62306149, 62407010, 62506109, 62502065), Jilin Province Philosophy and social science planning project/major project under Grant No.2023ZD15, Jilin Education Department Project under Grant No. (JJKH20250335KJ, JJKH20240208YJG), Social Science Research Project of the Jilin Provincial Department of Education under Grant No. JJKH20231252SK, Jilin Science and Technology Association under Grant No. QT202320, the Startup Foundation for Introducing Talent of NUIST under Grant No. 2023r045, and the Startup Foundation for Introducing Talent of DLMU No. 02502118.

## References

Adachi, M.; Hayakawa, S.; Jørgensen, M.; Wan, X.; Nguyen, V.; Oberhauser, H.; and Osborne, M. A. 2024. Adaptive batch sizes for active learning: A probabilistic numerics approach. In *International Conference on Artificial Intelligence and Statistics*, 496–504. PMLR.

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2021. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76.

Agboola, O. P.; and Yassin, Y. N. H. M. 2025. AI Applications in Education: Enhancing Human Creativity Through Collaborative Design. In *International Conference on Knowledge Management in Organizations*, 45–69. Springer.

Arthur, D.; and Vassilvitskii, S. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Ash, J.; Goel, S.; Krishnamurthy, A.; and Kakade, S. 2021. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34: 8927–8939.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Bhandari, U.; Chang, K.; and Neben, T. 2019. Understanding the impact of perceived visual aesthetics on user evaluations: An emotional perspective. *Information & management*, 56(1): 85–93.

Chauhan, R.; Ghanshala, K. K.; and Joshi, R. 2018. Convolutional neural network (CNN) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, 278–282. IEEE.

Chen, Z.; Yu, L.; Zhang, S.; Hu, S.; and Shen, C. 2024. Multiagent hierarchical deep reinforcement learning for operation optimization of grid-interactive efficient commercial buildings. *IEEE Transactions on Artificial Intelligence*, 5(8): 4280–4292.

Gaillochet, M.; Desrosiers, C.; and Lombaert, H. 2023. Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis*, 90: 102958.

Gao, M.; Zhang, Z.; Yu, G.; Arık, S. Ö.; Davis, L. S.; and Pfister, T. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part x 16*, 510–526. Springer.

Guo, M.; Lu, C.; Liu, Z.; Cheng, M.; and Hu, S. 2023. Visual attention network. *Computational Visual Media*, 9(4): 733–752.

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Y.; Lu, M.; Huang, W.; Yi, X.; and Li, T. 2025. Timefs: joint learning of tensorial incomplete multi-view unsupervised feature selection and missing-view imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17503–17510.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Jin, T.; Zhou, S.; Lang, X.; He, J.; and Wang, W. 2022. Combined effect of color and shape on cognitive performance. *Mathematical Problems in Engineering*, 2022(1): 3284313.

Juanzi, D.; et al. 2019. Appreciation and creation of children's English picture books. *Frontiers in Educational Research*, 2(2).

Kao, Y.; He, R.; and Huang, K. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3): 1482–1495.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Lin, K.; Zhao, R.; Xu, Z.; and Zhou, J. 2018. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1774–1783.

Liu, J.; Huang, X.; Zheng, J.; Liu, Y.; and Li, H. 2023. MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6252–6261.

Liu, K.; Fu, Y.; Wang, P.; Wu, L.; Bo, R.; and Li, X. 2019. Automating feature subspace exploration via multi-agent reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 207–215.

Liu, R.; Xie, R.; Yao, Z.; Fu, Y.; and Wang, D. 2025. Continuous optimization for feature selection with permutation-invariant embedding and policy-guided search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 1857–1866.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2015. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11): 2021–2034.

Mao, Q.; Hong, J.-C.; and Nguyen, H. B. N. 2023. Belief of aesthetic intelligence and attribution of failure in science abilities predict Chinese students' learning engagement in drawing a future science world. *Thinking Skills and Creativity*, 47: 101246.

Melo, L. C.; Tigas, P.; Abate, A.; and Gal, Y. 2024. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37: 118052–118085.

Qiming, X.; and Renganathan, S. 2024. Using Picturebooks as a Pedagogical Tool to Teach Drawing: A Case Study in a Preschool in China. *International Journal of Early Childhood*, 1–21.

Roth, D.; and Small, K. 2006. Margin-based active learning for structured output spaces. In *European conference on machine learning*, 413–424. Springer.

Rubashevskii, A.; Kotova, D.; and Panov, M. 2023. Scalable batch acquisition for deep bayesian active learning. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 739–747. SIAM.

Saran, A.; Yousefi, S.; Krishnamurthy, A.; Langford, J.; and Ash, J. T. 2023. Streaming active learning with deep neural networks. In *International Conference on Machine Learning*, 30005–30021. PMLR.

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Targ, S.; Almeida, D.; and Lyman, K. 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, D.; Wang, P.; Liu, K.; Zhou, Y.; Hughes, C. E.; and Fu, Y. 2021. Reinforced imitative graph representation learning for mobile user profiling: An adversarial training perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4410–4417.

Wang, W.; and Shen, J. 2017. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5): 2368–2378.

Wang, Z.; Chen, Z.; and Du, B. 2023. Active learning with co-auxiliary learning and multi-level diversity for image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 3899–3911.

Whiting, D. 2023. Admiration, appreciation, and aesthetic worth. *Australasian Journal of Philosophy*, 101(2): 375–389.

Xiao, M.; Wang, D.; Wu, M.; Qiao, Z.; Wang, P.; Liu, K.; Zhou, Y.; and Fu, Y. 2023. Traceable automatic feature transformation via cascading actor-critic agents. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 775–783. SIAM.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, 5571–5580. PMLR.

Yuan, Z.; and Zhang, J. 2016. Feature extraction and image retrieval based on AlexNet. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, volume 10033, 65–69. SPIE.

Zhang, J.; Miao, Y.; and Yu, J. 2021. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access*, 9: 77164–77187.

Zhang, Y.; Gao, Y.; Wang, D.; Zhou, Y.; He, J.; Sun, Z.; and Yin, M. 2025a. Multi-type MOOCs recommendation: leveraging deep multi-relational representation and hierarchical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13313–13321.

Zhang, Y.; Wang, M.; He, J.; Li, N.; Zhou, Y.; Huang, H.; Cai, D.; and Yin, M. 2024. AestheNet: Revolutionizing Aesthetic Perception Diagnosis in Education With Hybrid Deep Nets. *IEEE Transactions on Learning Technologies*.

Zhang, Y.; Wang, M.; He, J.; Zhou, Y.; Wu, H.; Sun, Z.; Zhang, Y.; and Yin, M. 2025b. Reinforcement Learning-Driven Optimization of Picture Book Paths for Aesthetic Perception Enhancement. *IEEE Transactions on Learning Technologies*.