AesPrompt: Zero-shot Image Aesthetics Assessment with Multi-Granularity Aesthetic Prompt Learning

Xiangfei Sheng, Leida Li, Pengfei Chen, Li Cai, Giuseppe Valenzise

AesPrompt: Zero-shot Image Aesthetics Assessment with Multi-Granularity Aesthetic Prompt Learning

Xiangfei Sheng, *Student Member, IEEE*, Leida Li, *Senior Member, IEEE*, Pengfei Chen, Li Cai, and Giuseppe Valenzise, *Senior Member, IEEE*

***Abstract*—Recent years have witnessed increasing interest towards image aesthetics assessment (IAA), which predicts the aesthetic appeal of images by simulating human perception. The state-of-the-art IAA methods, despite their significant advancements, typically rely heavily on time-consuming and labor-intensive human annotation of aesthetic scores. Furthermore, they are subject to the generalization challenge, which is highly desired in real-world applications. Motivated by this, zero-shot image aesthetics assessment (ZIAA) is investigated to achieve robust model generalization without relying on manual aesthetic annotations, which remains largely underexplored. Specifically, a novel aesthetic prompt learning framework for ZIAA, dubbed AesPrompt, is presented in this paper. The key insight of AesPrompt is to emulate the human aesthetic perception process for learning aesthetic-oriented prompts in a multi-granularity manner. First, we first develop a new pseudo aesthetic distribution generation paradigm based on multi-LLM ensemble. Then, external knowledge of multi-granularity prompts encompassing image themes, emotions, and aesthetics is acquired. Through learning the multi-granularity aesthetic-oriented prompts, the proposed method achieves better generalization and interpretability. Extensive experiments on five IAA benchmarks demonstrate that AesPrompt consistently outperforms the state-of-the-art ZIAA methods across diverse-sourced images, covering natural images, artistic images, and artificial intelligence-generated images. The source code is available at https://github.com/sxfly99/AesPrompt.**

***Index Terms*—Image Aesthetics Assessment, zero-shot learning, prompt learning, CLIP**

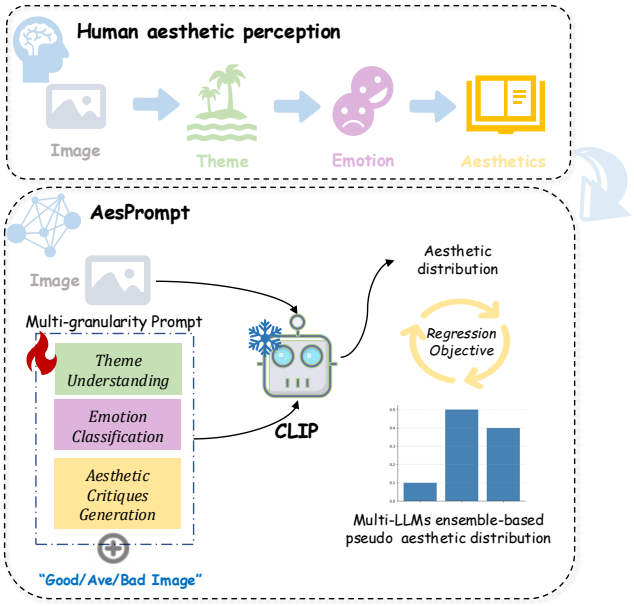


Fig. 1. Overview of our approach. Top: The human aesthetic perception process follows a hierarchical pattern from theme understanding to emotional response to aesthetic judgment. Bottom: Our AesPrompt framework emulates this process through multi-granularity prompt learning, integrating theme understanding, emotion classification, and aesthetic critiques generation. The framework leverages CLIP's vision-language capabilities and employs multi-LLM ensemble for generating reliable pseudo aesthetic distributions, enabling effective zero-shot aesthetics assessment.

I. Introduction

Image aesthetics assessment (IAA) is a fundamental task in image processing and multimedia, with the objective of predicting human perceptual judgments of image aesthetics in a computationally efficient manner. The significance of IAA has grown substantially in recent years, driven by its expanding applications across diverse domains. Traditional applications include image editing [1], photo enhancement [2], and automated content curation [3]. More recently, with the rapid development of generative AI, IAA has found new applications in emerging fields such as AI-generated image evaluation [4] and aesthetic-guided image synthesis [5]. This increasing relevance across both traditional and emerging domains has sparked intensive research interest in developing more effective and practical IAA methods.

Recent years have witnessed remarkable progress in IAA, with various approaches achieving impressive performance through different architectural designs, such as convolutional neural networks [6], [7], vision transformers [8], [9], and the emerging visual state space models [10], [11]. However, these state-of-the-art methods predominantly rely on substantial amounts of manually annotated aesthetic scores for training. This dependency poses two significant challenges. First, the annotation process is inherently labor-intensive and time-consuming, particularly given the subjective and abstract nature of aesthetics assessment. Second, the heavy reliance

This work was supported in part by the National Natural Science Foundation of China under Grants 62471349, 62171340, 62301378, in part by Fundamental Research Funds for the Central Universities under Grant QTZX25076, in part by the China Postdoctoral Science Foundation under Grant 2024M762553, and partly supported in part by the Fundamental Research Funds for the Central Universities, the Innovation Fund of Xidian University under Grant YJSJ24012. *(Corresponding author: Leida Li).*

Xiangfei Sheng, Pengfei Chen are with the School of Artificial Intelligence, Xidian University, China. (e-mails: xiangfeisheng@gmail.com; chenpengfei@xidian.edu.cn.)

Leida Li is with the School of Artificial Intelligence, Xidian University, China, and also with the School of Electronic and Information Engineering, Chongqing Three Gorges University, China. (e-mail: ldli@xidian.edu.cn.)

Li Cai is with the School of Electronic and Information Engineering, Chongqing Three Gorges University, China. (e-mail: 20040001@sanxiau.edu.cn,)

Giuseppe Valenzise is with the CNRS, CentraleSupelec, Laboratoire des Signaux et Systèmes, Université Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: giuseppe.valenzise@l2s.centralesupelec.fr).

on annotated data often leads to overfitting issues, potentially compromising the models' generalization capability in real-world scenarios.

In light of the above limitations, researchers have explored various low supervision-based methods for IAA, including semi-supervised [12], weakly-supervised [1], [13], and few-shot approaches [14]–[16]. Nevertheless, during the model training, these methods still require different amounts of aesthetic annotations from the target databases, which poses significant challenges in real-world deployment scenarios. First, the image types in target environments are often unknown a priori, requiring models to handle diverse visual domains (natural images, artistic images, and AI-generated images) with distinct aesthetic characteristics, where natural images demand evaluation of photographic principles, artistic images require assessment of creative expression, and AI-generated images necessitates examination of technical coherence. Second, data accessibility in practical applications is severely constrained by privacy concerns - obtaining user-generated images is inherently challenging, while collecting their aesthetic annotations becomes practically infeasible. In response to these challenges, zero-shot image aesthetics assessment (ZIAA) emerges as a promising paradigm that completely circumvents the dependency on manually annotated aesthetic scores during training [17], [18], while demonstrating robust generalization capabilities across diverse visual domains.

Despite the promising potential of ZIAA, research specifically designed for this task remains extremely limited. To the best of our knowledge, only KZIAA [18] has been specifically developed for ZIAA, which utilizes attribute-specific prompts and anchor images to transfer aesthetic knowledge. While some recent approaches [9], [17], [19] based on vision-language models (VLMs) like CLIP [20] and multimodal large language models (MLLMs) like LLaVA [21] exhibit certain zero-shot capabilities, they are not specifically designed for ZIAA task. These methods demonstrate limited zero-shot performance due to two fundamental limitations. First, existing approaches overlook the inherent multi-granularity nature of human aesthetic perception. As shown in Fig.1, recent psychological research [22] has revealed that aesthetic experience is a complex process involving multiple cognitive and emotional dimensions. Their findings demonstrate that humans process aesthetic information hierarchically: first understanding the thematic content, then experiencing emotional responses, and finally forming aesthetic judgments. Second, current methods predominantly adopt vision-language models as their foundation. However, the lack of reliable supervision signals has severely limited the rich aesthetic knowledge embedded in VLMs. Most existing methods rely on manually designed hard prompts (e.g., "Good/Average/Bad image") for zero-shot inference [9], [17].

To address these challenges, this paper introduce a novel approach, namely **AesPrompt**, to boost CLIP for accurate ZIAA across different domains. The key idea of AesPrompt is illustrated in Fig. 1, which is inspired by how humans utilize diversified contexts to perceive visual aesthetics [22]. To this end, we first establish a low-cost Pseudo Aesthetic Distribution Generation paradigm (PADG) based on multi-LLM ensemble, enabling learnable aesthetic-oriented prompts instead of hard prompts. Subsequently, we incorporate prompt learning into the ZIAA task by designing a Multi-granularity Aesthetic Prompt Learning (MAPL) strategy, which hierarchically categorizes the optimized aesthetic-oriented prompts into three distinct levels: theme-emotion-aesthetics. Finally, after training using the regression objective, the learned prompts are used for ZIAA. The contributions of this paper are as follows.

- We present AesPrompt, a psychologically-inspired aesthetic prompt learning framework for zero-shot image aesthetics assessment. By simulating the hierarchical property of human aesthetic perception, AesPrompt exhibits exceptional ZIAA performance across a wide range of image domains, including natural images, artistic images, and AI-generated images.
- We develop a simple yet universally-effective Pseudo Aesthetic Distribution Generation paradigm based on multi-LLM ensemble. By utilizing the generated aesthetic distribution label, we significantly simplify the prompt design and optimize prompt learning for ZIAA task through a regression objective without requiring manual aesthetic annotations.
- We design a Multi-granularity Aesthetic Prompt Learning strategy, which learns aesthetic-oriented prompts in a multi-granularity manner, featuring interpretability and zero-shot generalization.

II. Related Works

A. Image Aesthetics Assessment

IAA has evolved significantly over the past decades [23]. Early approaches primarily relied on hand-crafted features designed to capture photographic rules and aesthetic principles. Pioneering works [24], [25] attempted to extract low-level visual features such as color distribution, composition rules, and lighting conditions. Subsequently, researchers incorporated more sophisticated features by considering subject regions [26] and photographic composition guidelines [27], [28]. While these traditional methods laid the foundation for computational aesthetics, their performance was limited by the hand-designed nature of the features.

The advent of deep learning has revolutionized IAA research, leading to two main paradigms: single-modal and multi-modal approaches. In single-modal approaches, early attempts focused on patch-based strategies to capture both local details and global composition. Lu *et al.* [29] proposed a double-column network architecture to process global and local views simultaneously. This idea was further extended to multi-patch aggregation [30] and adaptive layout-aware frameworks [31], [32] to better preserve image composition. Recent advances have explored multi-task learning frameworks to leverage aesthetic attributes [33] and personality traits [34] for more comprehensive assessment. Graph Convolutional Networks-based methods have also emerged, effectively modeling the relationships between image regions and visual attributes for enhanced aesthetic reasoning [35], [36]. Multi-modal approaches have gained increasing attention by incorporating textual modality beyond visual features [37].

Zhou *et al.* [38] pioneered the use of textual information by proposing a joint image-text representation using Deep Boltzmann Machine. Zhang *et al.* [39] further advanced this direction by developing a multimodal recurrent attention network that combines visual attention mechanisms with textual feature learning. Recent works have explored more sophisticated multi-modal architectures. For example, Niu *et al.* [40] leveraged comment-guided semantics for enhanced aesthetic understanding, while Li *et al.* [41] proposed an attribute-assisted memory network to model fine-grained interactions between visual and textual modalities. These multi-modal methods have demonstrated that incorporating textual information can provide valuable semantic cues and improve aesthetics assessment accuracy.

Despite the impressive progress, most of the existing IAA methods are based on supervised learning, where models are explicitly trained using labeled data obtained through labor-intensive user studies. Further, this paradigm makes IAA models easy to overfit and leads to unsatisfied generalization performance. To address these limitations, this paper investigates zero-shot image aesthetics assessment (ZIAA), which aims to evaluate image aesthetics without requiring any ground-truth aesthetic scores during training while maintaining robust generalization performance.

B. Zero-shot Image Aesthetics Assessment

ZIAA aims to evaluate image aesthetics without utilizing any ground-truth aesthetic scores during training. Despite its practical significance, research specifically designed for ZIAA remains extremely limited. To the best of our knowledge, KZIAA [18] is the only model specifically designed for the ZIAA task, where a knowledge-driven ZIAA model was proposed by leveraging knowledge through attribute-specific prompt tuning. By computing aesthetic scores based on selected anchor images without, they demonstrated the feasibility of zero-shot aesthetics assessment.

Recent vision-language models (VLMs) and multimodal large language models (MLLMs) exhibit zero-shot assessment capabilities due to their model architectures, although they are not specifically designed for ZIAA. For instance, Wang *et al.* [42] first demonstrated the potential of leveraging CLIP's vision-language knowledge for zero-shot quality assessment by introducing an antonym prompt pairing strategy. Sheng *et al.* [9] introduced a multi-attribute contrastive learning framework to bridge the domain gap between general visual understanding and aesthetics assessment. Ke *et al.* proposed VILA [17], which leveraged user comments through vision-language pretraining to learn rich aesthetic semantics. Despite these advances, VLM-based methods often rely on hand-crafted prompts, which may not fully exploit the aesthetic knowledge embedded in VLMs. More recently, some MLLM-based aesthetic expert models have been proposed [19], [43], [44], showing promising results in providing interpretable aesthetic critiques. However, they often struggle with accurate score prediction since MLLMs are primarily designed for text generation rather than numerical assessment.

In this paper, we propose AesPrompt, a novel prompt learning framework that fundamentally differs from existing ZIAA approaches by introducing learnable aesthetic-oriented prompts optimized through pseudo aesthetic distributions, rather than relying on hand-crafted prompts. This design enables more effective extraction of aesthetic knowledge from VLMs while maintaining interpretability through a multi-granularity learning strategy.

C. Prompt Learning

In the era of large-scale models, prompt learning has emerged as a powerful technique for adapting pretrained models to downstream tasks. The core idea is to introduce learnable continuous vectors, which are optimized to modulate the model's attention mechanisms and guide its predictions towards the target task. These learnable prompts can be either static [45] or dynamically conditioned on input features [46], [47], demonstrating exceptional performance across diverse tasks in both zero-shot and few-shot settings. For instance, Zhou *et al.* [45] proposed Context Optimization (CoOp), which involved adapting CLIP-like VLMs for downstream task by incorporating learnable vectors. To enhance the ability to generalize to unseen classes, Zhou *et al.* further introduced Conditional Context Optimization (CoCoOp) [46], which leveraged dynamic prompts by integrating visual features. More recently, Khattak *et al.* [47] proposed Multi-modal Prompt Learning (MaPLe) as a means of enhancing the alignment between vision and language representations. The majority of existing prompt learning methods are *task-independent* and when applied to the ZIAA task, they only yield sub-optimal performance.

III. Proposed AesPrompt

In this section, we elaborate on the proposed AesPrompt framework for zero-shot image aesthetics assessment. As illustrated in Fig. 2, AesPrompt consists of three main components. First, the Pseudo Aesthetic Distribution Generation module leverages multi-LLM ensemble to generate reliable pseudo labels, enabling prompt learning without manual annotations. Second, the Multi-Granularity Aesthetic Prompt Learning module learns aesthetic-oriented prompts hierarchically by simulating human aesthetic perception process through theme understanding, emotion recognition, and aesthetic critique generation. Third, the Multi-Attribute Prompts Ensemble module integrates the well-learned prompts from different aesthetic attributes for comprehensive assessment during inference. In the following subsections, we introduce the three stages in further detail.

A. Preliminaries

Problem Definition. ZIAA aims to evaluate the aesthetic quality of images **without utilizing any manually annotated aesthetic scores during training**. Unlike conventional zero-shot learning tasks where categories are discrete and well-defined, aesthetic assessment involves inherently subjective and continuous judgments that vary across different observers and contexts. To clarify the scope of "zero-shot" in this context, it is important to distinguish between different types of annotations:

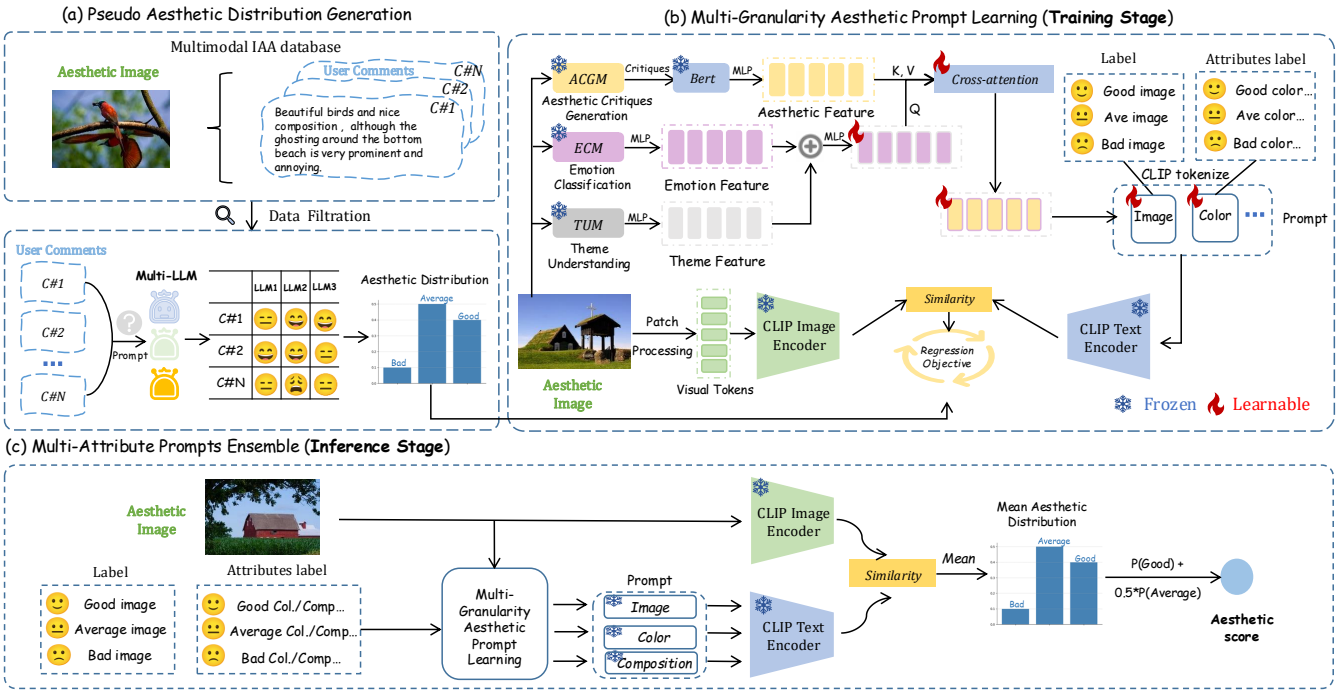


Fig. 2. Overview of the proposed AesPrompt framework. (a) Pseudo Aesthetic Distribution Generation (PADG) leverages multi-LLM ensemble to generate reliable pseudo labels from aesthetic comments. (b) Multi-Granularity Aesthetic Prompt Learning (MAPL) incorporates theme, emotion, and aesthetic knowledge through a hierarchical learning strategy. (c) Multi-Attribute Prompts Ensemble (MAPE) integrates different aesthetic attributes for comprehensive assessment during inference. The framework enables effective zero-shot aesthetics assessment by simulating human perception process.

1) *Aesthetic scores*: Numerical ratings assigned by human annotators that directly quantify the aesthetic quality of images. These explicit annotations are *not used at any stage* in ZIAA methods.

2) *Aesthetic comments*: Textual descriptions or critiques of images that contain implicit aesthetic judgments. These textual annotations *may be leveraged* to extract aesthetic knowledge, but they do not provide direct numerical ratings.

In accordance with established practices in recent ZIAA research [9], [17], [18], our definition permits the use of textual comments while strictly prohibiting the use of numerical aesthetic scores during training. This approach preserves the "zero-shot" nature of the task while enabling the extraction of rich aesthetic knowledge embedded in natural language descriptions. Formally, let $\mathcal{X}$ denote the image space and $\mathcal{S} = [0, 1]$ represent the normalized aesthetic score space. Given a training set $\mathcal{D}_{train} = \{x_i, \{c_i^j\}_{j=1}^{J_i}\}_{i=1}^{N}$ where $x_i \in \mathcal{X}$ and $\{c_i^j\}_{j=1}^{J_i}$ represents a set of textual comments associated with image x_i (but no aesthetic scores), and a test set $\mathcal{D}_{test} = \{(x_j, \{s_j^k\}_{k=1}^{K})\}_{j=1}^{M}$ where $\{s_j^k\}_{k=1}^{K}$ represents multiple human aesthetic judgments for image x_j, ZIAA aims to learn a mapping function $f : \mathcal{X} \rightarrow \mathcal{S}$ that can predict aesthetic scores for unseen images.

The key challenges lie in: (1) the zero-shot constraint where no aesthetic scores are available during training, and (2) the need to capture the subjective and multi-faceted nature of aesthetic perception, which requires the model to understand and transfer complex aesthetic knowledge across different visual domains and aesthetic criteria.

B. Pseudo Aesthetic Distribution Generation

A fundamental challenge in developing learnable prompts for ZIAA is **the lack of reliable supervision signals for prompt optimization.** To address this challenge, we propose a novel PADG paradigm that leverages the collective intelligence of multiple Large Language Models (LLMs), which is illustrated in Fig. 2(a). Specifically, the proposed PADG paradigm comprises two steps, including data filtration and multi-LLM based distribution generation.

For a fair comparison with existing ZIAA methods [9], [17], we utilize the AVA-comments database [38] for prompt learning in implementation. To address potential concerns regarding **data leakage**, we maintain the same split in the AVA-Comments database as in the AVA database. Moreover, our model exclusively utilizes the data solely from the training set of AVA-comments.

Data Filtration. The AVA-comments dataset [38] contains user-provided aesthetic comments collected from DPChallenge, a photography website where users can rate and comment on images. Intuitively, these comments often contain valuable aesthetic judgments. Motivated by this fact, we try to mine aesthetic level information from aesthetic comments. To ensure the quality of pseudo-labels, we design a two-stage filtration strategy: First, we conduct semantic analysis to eliminate comments that lack meaningful aesthetic content or are too brief. Second, we implement a density-based filtering mechanism that retains only images with at least three substantive aesthetic comments. This process results in a refined dataset of approximately 125K samples from the original 230K

AVA training set, maintaining consistency with existing ZIAA methods [9], [17] to prevent data leakage.

Entropy-guided Multi-LLM Distribution Ensemble. As shown in Fig. 2, we propose an entropy-guided ensemble approach that leverages multiple state-of-the-art LLMs to generate more reliable aesthetic distributions. Specifically, we employ three LLMs: Llama3-70B [48] (represented with a blue icon), Qwen 1.5-32B [49] (depicted with a green icon), and Mistral-7B [50] (illustrated with an orange icon). This diverse ensemble allows us to capture different aspects of aesthetic understanding while mitigating individual model biases. Following the standard two-stage prompting paradigm in MLLMs, we first define the system prompt to establish the LLM's role, followed by a user prompt that specifies the task instruction:

***#System**: You are an expert in the aesthetic critique of images.*

***#User**: You are given aesthetic comments of the same image from different reviewers. For each aesthetic comment, please give the aesthetic classification result of the comment in a single word in 'Good, Average, Bad'. Organize the output a list in JSON format: {'c_[x]': result} and when you respond, please only output the json, no other words are needed. The aesthetic comments are: [Comments].*

For each image x, each LLM l_i $(i = 1, 2, ..., m)$ processes the associated comments and outputs a probability distribution $p_i \in \mathbb{R}^3$ over three aesthetic levels (Good, Average, Bad). We choose this three-level categorization as it aligns with common aesthetic rating practices while maintaining reliable LLM predictions. To optimally combine these distributions from different LLMs, we propose an entropy-weighted ensemble strategy:

$$H(p_i) = -\sum_k p_i(k) \log p_i(k), \tag{1}$$

$$p_f = \sum_{i=1}^{m} \frac{p_i / H(p_i)}{\sum_{j=1}^{m} 1/H(p_j)}, \tag{2}$$

where $H(p_i)$ represents the entropy of distribution p_i from the i-th LLM, and $k \in \{Good, Average, Bad\}$ indexes the three aesthetic levels. This approach assigns higher weights to more confident predictions (lower entropy) in the final ensemble. Furthermore, our framework can be readily extended to generate pseudo-distributions for specific aesthetic attributes by modifying the LLM instruction template: "*For each aesthetic comment, please give the* `attribute` *classification result of the critique in a single word in 'Good, Average, Bad'*", where `attribute` could be color, composition, etc. These attribute-specific distributions enable the training of specialized aesthetic prompts.

C. Multi-Granularity Aesthetic Prompt Learning

When applying prompt learning to ZIAA task, a straightforward approach is to introduce trainable parameters as context vectors for vision-language models. As illustrated in Fig. 3(a), existing general-purpose prompt learning methods like CoOp [45] and CoCoOp [46] typically rely on random initialization of these prompts. However, recent studies have shown that such random initialization can lead to unstable optimization trajectories in specific vision-language tasks, particularly when the target task requires fine-grained understanding of domain-specific concepts [46], [47]. Our empirical analysis (detailed in Section IV.C) further demonstrates this phenomenon in the context of aesthetics assessment, where randomly initialized prompts exhibit high variance in their optimization trajectories.

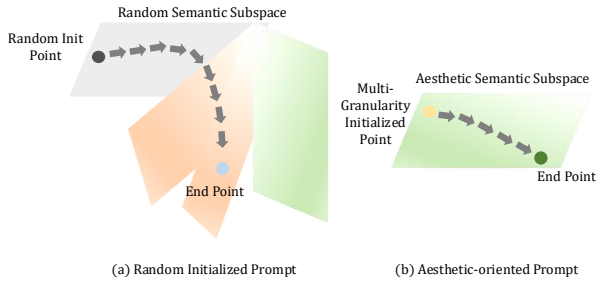


Fig. 3. Visualization of prompt optimization trajectories in semantic space. (a) Random initialization leads to optimization through multiple semantic subspaces (colored regions representing different semantic concepts), resulting in unstable convergence. (b) Our aesthetic-oriented initialization, guided by multi-granularity knowledge, provides a more direct optimization path within the aesthetic-relevant semantic subspace (green region).

To address this limitation, we propose MAPL, a framework specifically designed for aesthetics assessment. As shown in Fig. 3(b), instead of random initialization, MAPL incorporates aesthetic domain knowledge to guide the prompt optimization process. Inspired by the success of knowledge-guided initialization in vision-language tasks [46], we hypothesize that such domain-specific initialization can provide more stable optimization trajectories within the aesthetic-relevant semantic subspace. Our experimental results (Section IV.C) validate this hypothesis, demonstrating that MAPL achieves more consistent convergence and superior performance compared to random initialization baselines.

1) Aesthetic Prior Learning: Motivated by psychological research [22] that humans perceive image aesthetics through a hierarchical process (*theme-emotion-aesthetics*), we first construct a set of aesthetic prior models to capture different granularities of aesthetic perception. As illustrated in Fig. 4, we design three complementary prior models: a theme understanding model for capturing high-level semantic concepts, an emotion classification model for recognizing affective characteristics, and an aesthetic critiques generation model for providing fine-grained aesthetics assessments. This multi-granularity prior knowledge serves as a strong initialization for our prompt learning process, helping to stabilize the optimization of prompt parameters and ensure more effective aesthetic knowledge transfer.

Theme Understanding Model (TUM). To encode high-level semantic concepts, we develop a theme understanding model $\mathcal{M}_{TUM}$ trained on the SPAQ dataset [51]. The model learns to classify images into nine theme categories that are widely recognized in photographic aesthetics [51]: animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others. These categories were chosen as they

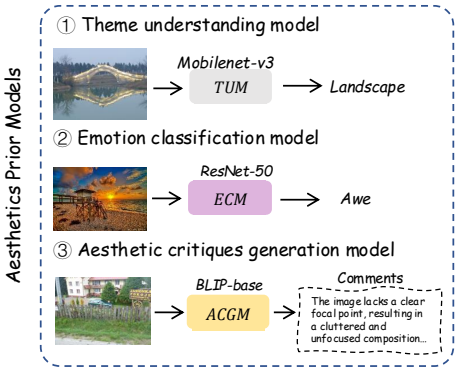


Fig. 4. Illustration of our aesthetic prior models operating at different granularities. The Theme Understanding Model captures high-level semantic concepts of the image content. The Emotion Classification Model recognizes affective characteristics conveyed by the image. The Aesthetic Critiques Generation Model provides fine-grained aesthetics assessments. These complementary models collectively establish a comprehensive foundation for aesthetic prompt learning.

represent distinct photographic genres with well-established aesthetic principles - for instance, landscape photography emphasizes composition and lighting, while portrait (human) photography focuses on subject presentation and emotional connection. For an input image x, the theme representation is computed as:

$$F_t = \Phi_{\mathcal{T}}(\mathcal{M}_{TUM}(x;\Theta_{TUM})) \in \mathbb{R}^{d_t \times h_t}, \tag{3}$$

where $\Phi_{\mathcal{T}}$ denotes the feature extraction operation, Θ_{TUM} represents learnable parameters, and d_t, h_t are feature dimensions.

Emotion Classification Model (ECM). To capture affective characteristics, we construct an emotion classifier $\mathcal{M}_{ECM}$ using the Emoset-118K dataset [52]. Following established research in aesthetic psychology [22], we consider eight fundamental emotions that have been shown to influence aesthetic perception: amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. These emotions have distinct correlations with aesthetic quality - for instance, images evoking awe often exhibit high aesthetic value through their grandeur or sublime qualities, while those inducing contentment typically demonstrate balanced and harmonious compositions [22]. The model is built upon CLIP's ResNet50 backbone, and the emotion features are extracted through:

$$F_e = \Phi_{\mathcal{E}}(\mathcal{M}_{ECM}(x;\Theta_{ECM})) \in \mathbb{R}^{d_e \times h_e}, \tag{4}$$

where $\Phi_{\mathcal{E}}$ is the emotion feature extractor with parameters Θ_{ECM}.

Aesthetic Critiques Generation Model (ACGM). To extract fine-grained aesthetic knowledge without relying on numerical scores, we propose a critique generation model based on BLIP [53]. Through our analysis of the AVA-Comments dataset, we observe that a significant portion of user comments consists of noisy, uninformative expressions. To address this issue, we introduce a LLM-guided refinement strategy that transforms these raw comments into structured, informative aesthetic descriptions (detailed qualitative results are presented in Section IV.E). Based on this refined training data, for an input image x, the ACGM generates aesthetic descriptions through an autoregressive process:

$$c = \mathcal{M}_{ACGM}(x;\Theta_{ACGM}) \in \mathcal{V}^L, \tag{5}$$

where $\mathcal{V}^L$ represents the space of possible text sequences with vocabulary $\mathcal{V}$ and length L. The textual aesthetic features F_a are then extracted through:

$$F_a = \Phi_{\mathcal{A}}(BERT(c)) \in \mathbb{R}^{d_a \times h_a}, \tag{6}$$

where $\Phi_{\mathcal{A}}$ represents a feature transformation function that maps BERT's contextual embeddings to aesthetic-specific representations.

2) Aesthetic-oriented Prompt Learning: To effectively model the hierarchical nature of aesthetic perception, we propose a multi-stage prompt learning strategy that progressively integrates theme, emotion, and aesthetic information. First, we design a Theme-Emotion Fusion Module (TEFM) to capture the intricate relationships between thematic and emotional features:

$$F_{te} = F_t + \mathcal{T}(\xi(F_t) \oplus \xi(F_e)), \tag{7}$$

where $\mathcal{T} : \mathbb{R}^{d_t+d_e} \to \mathbb{R}^{d_t}$ is a learnable transformation function implemented as an MLP, $\xi(\cdot)$ denotes batch normalization, and $\oplus$ represents feature concatenation.

Subsequently, we employ a Cross-modal Attention Module (CAM) to establish dynamic relationships between the fused theme-emotion features and aesthetic critiques. The attention mechanism is formulated as:

$$\begin{aligned} Q &= W_q F_{te}, K = W_k F_a, V = W_v F_a \\ A &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \\ F_{aes} &= [AV \oplus F_l^1; AV \oplus F_l^2; AV \oplus F_l^3], \end{aligned} \tag{8}$$

where $W_q \in \mathbb{R}^{d_k \times d_t}$, $W_k, W_v \in \mathbb{R}^{d_k \times d_a}$ are learnable projection matrices, and $F_l^i \in \mathbb{R}^{d_l}$ represents the label token embeddings for different aesthetic levels.

The aesthetic distribution is then computed through cosine similarity:

$$s = \text{softmax}(\{\cos(F_{aes}^i, F_{vis})\}_{i=1}^3), \tag{9}$$

where F_{vis} denotes the features of the input images from CLIP image encoder. Finally, the Earth Mover's Distance (EMD) loss is employed to optimize the learnable aesthetic-oriented prompts by measuring the distance between the predicted aesthetic distribution s and the pseudo aesthetic distribution $\hat{s}$, which is defined as

$$\mathcal{L}_{reg} = \left(\frac{1}{N} \sum_{k=1}^{N} |CDF_s(k) - CDF_{\hat{s}}(k)| \right)^{\frac{1}{2}}, \tag{10}$$

where N represents the number of images, and CDF denotes the cumulative distribution function.

D. Multi-Attributes Prompts Ensemble

To achieve more comprehensive aesthetics assessment, we leverage our PADG paradigm to generate not only overall aesthetic distributions but also attribute-specific pseudo distributions simultaneously. Specifically, by modifying the LLM instruction template to focus on different aesthetic attributes, we obtain a set of attribute-specific pseudo labels for prompt learning. In our implementation, we focus on four key aesthetic attributes that collectively cover the major dimensions of visual aesthetics, including composition, color, light and content. These attributes were selected based on their prevalence in aesthetic literature [54], [55] and their ability to comprehensively capture different aspects of visual aesthetics. For each attribute a_i, we learn a specialized prompt following the same procedure described in Section III.C, but with corresponding attribute-specific pseudo distributions.

Let $\mathcal{A} = \{a_1, ..., a_M\}$ denote a set of M aesthetic attributes. For each attribute a_i, we learn a specialized prompt following the same procedure described in Section III.C, but with corresponding attribute-specific pseudo distributions. During inference, these well-learned attribute prompts can be efficiently embedded into the CLIP model with minimal computational overhead. Given an input image x, we first obtain a set of predicted distributions:

$$\begin{aligned} &S = \{s^{img}, s^{a_1}, ..., s^{a_M}\} \in \mathbb{R}^{(M+1)\times 3}, \\ &\bar{s}_k = \frac{1}{M+1}\sum_{i=1}^{M+1} s_k^i, \quad k \in \{good, average, bad\}, \end{aligned} \tag{11}$$

where s^{img} represents the overall aesthetic distribution and s^{a_i} denotes the distribution for attribute a_i.

The final aesthetic score is then computed through a weighted combination that reflects the natural aesthetic rating scale:

$$score = 1.0 \cdot \bar{s}_{good} + 0.5 \cdot \bar{s}_{average} + 0.0 \cdot \bar{s}_{bad}, \tag{12}$$

where the weights (1.0, 0.5, 0.0) are designed to map the categorical distributions to a normalized score range [0,1].

IV. Experiments

A. Databases

To evaluate the effectiveness of our proposed method, we conducted extensive experiments on five public benchmark IAA databases:

AVA database [56] is the most popular database in the field of IAA task, which encompasses a vast collection of over 250,000 images sourced from DPChallenge. In accordance with the official split, this paper utilize 12,776 images for testing.

AADB database [55] is a collection designed specifically for attribute prediction and aesthetic score regression tasks. The dataset consists of 10,000 images collected from Flickr, with 8,500 images allocated for training, 500 images for validation, and 1,000 images for testing.

TAD66K database [57] comprises 66,000 images encompassing 47 prominent themes, with each image meticulously annotated by over 1200 individuals based on dedicated theme evaluation criteria. Following the official split, there are 14,079 images for testing.

APDD database [58] is a comprehensive collection of 4,985 artistic images, encompassing 24 distinct artistic categories and 10 different aesthetic attributes. Following the official split, there are specifically designated 498 images for testing.

SAC database [59] comprises a dataset of over 238,000 synthetic images generated using AI models like Stable Diffusion. These images have been rated by users on a scale of 1 to 10 in terms of their aesthetic value. For testing purposes, there are a total of 29,275 images available.

B. Experimental Settings

We use a pretrained ViT-B/16 CLIP model as our backbone. The resolution of the input image is set to 224 × 224. During the vision-language aesthetics pretraining, we conducted a 20-epoch fine-tuning of the CLIP model on AVA-Comments. Initially, the learning rate is set to 1e-4 with the batch size of 64. In the prompt learning stage, model parameters of CLIP are all frozen, and the length of learnable prompts is set to 16. We train the learnable prompts for 20 epoch, and learning rate of 1e-4. We use the linear decay schedule in our experiments. And our model is optimized using the AdamW. All the experiments are conducted on a machine equipped with two NVIDIA GeForce RTX 4090 GPUs. Considering that most of the existing ZIAA methods solely report their score regression performance, we employ Pearson linear correlation coefficient (PLCC) and Spearman's Rank Correlation (SRCC) as evaluation metrics.

C. Performance Comparison

1) Comparison with state-of-the-arts: As shown in Tables I and II, we evaluate the zero-shot performance of existing IAA methods on five databases, which cover diverse image sources including natural images, artistic images, and artificial intelligence-generated images.

Natural Image Databases. On natural image databases (AVA [56], AADB [55], and TAD66K [57]), as shown in Tables I, AesPrompt achieves state-of-the-art performance. Specifically, compared with MLLM-based methods, AesPrompt significantly outperforms AesExpert [19] on TAD66K, improving the PLCC from 0.192 to 0.432, even though AesExpert was trained with TAD66K data. Moreover, our method surpasses UNIAA [44] on both AVA and TAD66K without requiring a larger model or additional training data. Among VLM-based approaches, despite sharing similar backbone architecture, AesPrompt substantially outperforms previous methods, achieving a PLCC of 0.727 on AVA database, which represents a 6.3% improvement over VILA [17]. This significant improvement demonstrates the effectiveness of our aesthetic-oriented prompt learning in extracting aesthetic knowledge from VLMs.

Artistic Image Database. On the APDD database, which contains artistic images, AesPrompt exhibits very encouraging performance with a PLCC of 0.661, significantly surpassing all baselines. Notably, our method demonstrates stronger generalization ability compared to AesExpert [19], despite AesExpert's advantage of being trained with artistic images. This

TABLE I
PERFORMANCE COMPARISION OF THE PROPOSED AESPROMPT WITH THE STATE-OF-THE-ART ON THREE NATURAL IMAGE IAA DATABASES. THE BEST AND SECOND-BEST PERFORMANCES FROM EACH DATABASE ARE HIGHLIGHTED AND UNDERLINED. ENS.: ENSEMBLE PROMPTS.

Method	Ens.	Seen IAA data	AVA		AADB		TAD66K	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
MLLM-based methods								
AesExpert [19]	✗	AADB, TAD66K	0.394	0.380	0.558	**0.587**	0.192	0.173
UNIAA [44]	✓	AVA, AADB	0.704	0.713	-	-	<u>0.425</u>	<u>0.411</u>
VLM-based methods								
CLIP-IQA [42]	✗	-	0.341	0.317	0.403	0.359	0.184	0.164
KZIAA [18]	✓	AVA	0.454	0.446	-	-	-	-
AesCLIP [9]	✓	AVA	0.647	0.637	0.537	0.529	0.383	0.370
VILA [17]	✓	AVA	0.664	0.658	0.501	0.476	0.372	0.350
AesPrompt (Single Prompts)	✗	AVA	<u>0.716</u>	<u>0.714</u>	<u>0.561</u>	0.545	0.422	0.410
AesPrompt (Ensemble Prompts)	✓	AVA	**0.727**	**0.726**	**0.572**	<u>0.556</u>	**0.432**	**0.418**

TABLE II
PERFORMANCE COMPARISION OF THE PROPOSED AESPROMPT WITH THE STATE-OF-THE-ART ON APDD AND SAC DATABASES.

Method	APDD		SAC	
	PLCC	SRCC	PLCC	SRCC
AesExpert [19]	0.311	0.330	0.244	0.263
CLIP-IQA [42]	0.304	0.278	0.262	0.244
AesCLIP [9]	0.563	0.507	0.292	0.291
VILA [17]	0.541	0.542	0.288	0.289
AesPrompt (SP)	<u>0.631</u>	<u>0.578</u>	<u>0.319</u>	<u>0.334</u>
AesPrompt (EP)	**0.661**	**0.582**	**0.324**	**0.341**

substantial performance gap suggests potential limitations of MLLM-based methods in accurate aesthetic score prediction, despite their strong language understanding capabilities.

AI-Generated Image Database. For the SAC [59] database containing AI-generated images, AesPrompt again demonstrates superior performance, outperforming the second-best method AesCLIP [9] by 3.2% and 5.0% respectively. This consistent excellence across diverse image sources, particularly on challenging out-of-distribution samples like AI-generated images, validates the strong generalization capability of our approach.

Despite being trained primarily on natural images, our method demonstrates superior performance on both artistic images (APDD) and AI-generated images (SAC). This robust generalization can be attributed to three key factors: (1) **Multi-granularity knowledge transfer**: By decomposing aesthetic perception into hierarchical components (theme, emotion, and aesthetics), our model captures domain-agnostic principles that transfer effectively across visual domains; (2) **Prompt learning strategy**: Unlike traditional supervised approaches that may overfit to domain-specific features, our prompt learning mechanism enables the model to learn flexible aesthetic representations that adapt to diverse visual contexts; (3) **Psychological foundation**: Our approach, grounded in psychological research on human aesthetic perception, inherently models the domain-invariant cognitive processes that humans employ when evaluating aesthetics across different types of visual media. These factors collectively enable AesPrompt to bridge the domain gap between natural, artistic, and AI-generated images more effectively than previous methods.

It is important to emphasize that all VLM-based ZIAA methods in our comparison utilize the same AVA-comments dataset during their respective training stages. Although each method implements different technical approaches, with AesCLIP [9] applying attribute-oriented contrastive learning, VILA [17] leveraging vision-language alignment with comments, and our method generating pseudo aesthetic distributions, they all share the same underlying data source. This consistency ensures our experimental comparisons remain fair and methodologically sound. Additionally, all these methods maintain their zero-shot nature by exclusively utilizing textual comments for learning aesthetic representations without accessing any ground truth aesthetic scores during training.

2) Comparison of prompt learning methods: To demonstrate the efficacy of the proposed multi-granularity aesthetic prompt learning, we conduct a comparative analysis with popular task-independent prompt learning methods, which is listed in Table III. We first establish two baselines: vanilla CLIP without any pretraining, and CLIP with vision-language aesthetic pretraining. The results show that aesthetic pretraining substantially improves performance, more than doubling the PLCC/SRCC metric on both databases.

When incorporating learnable prompts, general-purpose prompt learning methods like CoOp [45] and CoCoOp [46] show further improvements over the pretrained baseline. However, on the artistic image database APDD, both methods show performance degradation compared to the pretrained baseline (SRCC dropping from 0.578 to 0.552 and 0.481 respectively). This degradation might be attributed to the domain gap between natural and artistic images - these task-independent prompt learning methods, primarily designed for natural image classification, may struggle to capture the

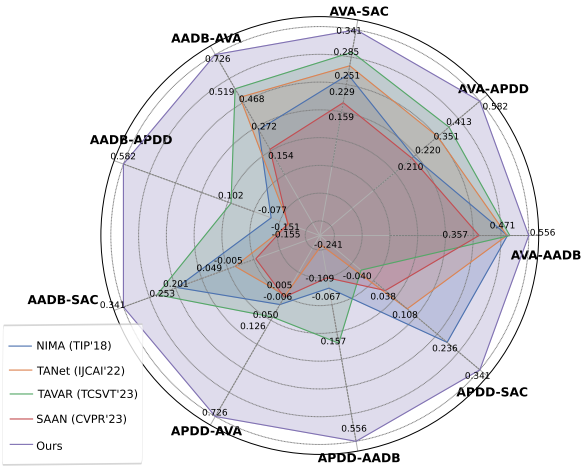


Fig. 5. Performance comparison of the proposed AesPrompt (zero-shot) with the state-of-the-art supervised IAA methods (cross-dataset). Metric is SRCC. Each dimension in the radar plot is scaled to the maximum correlation achieved on that dataset.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT PROMPT LEARNING METHODS ON AVA AND APDD DATABASE (SINGLE PROMPT).

Method	AVA		APDD	
	PLCC	SRCC	PLCC	SRCC
Vanilla CLIP	0.254	0.257	0.129	0.116
w Vision-Language Aesthetic Pretraining				
CLIP (baseline)	0.575	0.574	0.597	0.578
w proposed PADG & Learnable Prompt				
CoOp [45]	0.682	0.680	0.553	0.552
CoCoOp [46]	0.697	0.693	0.557	0.481
AesPrompt	**0.716**	**0.713**	**0.631**	**0.587**

unique aesthetic characteristics of artistic images, leading to sub-optimal prompt optimization. In contrast, our proposed AesPrompt demonstrates robust performance across both databases, achieving SRCC of 0.713 on AVA and 0.578 on APDD. The consistent improvement, especially on the challenging artistic database, validates the effectiveness of our aesthetic-oriented prompt learning strategy. By incorporating domain-specific aesthetic knowledge through multi-granularity prompt learning, our method could achieve better generalization performance across diverse domains.

3) Comparison with supervised IAA methods: To further demonstrate the strong generalization capability of AesPrompt, we conduct comprehensive comparisons between our zero-shot approach and existing supervised IAA methods under two distinct evaluation settings.

Cross-dataset Setting. To thoroughly evaluate model generalization, we conduct extensive cross-dataset experiments comparing AesPrompt with state-of-the-art supervised IAA methods, where supervised methods are trained on one dataset and tested on another, while our method maintains its zero-shot setting. As illustrated in Fig. 5, we analyze the performance across different database pairs, where "A-B" denotes training on database A and testing on database B. Our analysis reveals several key findings: 1) Natural Image Transfer: When transferring between natural image databases (AVA-AADB and AADB-AVA), AesPrompt demonstrates superior generalization ability, achieving SRCC of 0.556 and 0.726 respectively. 2) Artistic Domain Adaptation: The generalization gap becomes more pronounced when evaluating on the artistic database APDD. While supervised methods struggle with this domain shift, AesPrompt maintains strong performance. 3) AI-Generated Images: Most notably, on the challenging SAC database containing AI-generated images, AesPrompt achieves consistent performance across different training sources.

TABLE IV
PERFORMANCE COMPARISON WITH SUPERVISED IAA METHODS ON AVA AND TAD66K DATASETS.

Method	AVA		TAD66K	
	PLCC	SRCC	PLCC	SRCC
NIMA [6]	0.636	0.612	0.405	0.390
A-Lamp [31]	0.671	0.666	0.422	0.421
Zeng *et al.* [60]	0.720	0.719	0.441	0.433
HLA-GCN [35]	0.687	0.665	<u>0.493</u>	**0.486**
BIAA [61]	0.668	0.651	0.431	0.417
AesMamba [11]	**0.760**	**0.751**	**0.503**	<u>0.475</u>
AesPrompt	<u>0.727</u>	<u>0.726</u>	0.431	0.418

TABLE V
PERFORMANCE COMPARISON WITH SUPERVISED IAA METHODS ON APDD DATASET.

Method	APDD	
	PLCC	SRCC
TANet [57]	0.603	0.563
SAAN [62]	**0.672**	**0.619**
AANSPS [58]	-	<u>0.609</u>
AesPrompt	<u>0.661</u>	0.582

The significant performance degradation of supervised IAA methods in cross-dataset evaluation deserves further consideration. As shown in Fig. 5, some supervised methods even exhibit negative correlations when tested across different datasets. This phenomenon can be primarily attributed to two factors: (1) Dataset-specific overfitting: With the limited scale of aesthetic datasets, supervised models tend to learn dataset-specific characteristics rather than universal aesthetic principles; (2) Visual domain gaps: The substantial differences between natural photographs, artistic images, and AI-generated content create a challenging generalization problem for models trained on a single domain. In contrast, our AesPrompt demonstrates robust cross-domain performance by leveraging prompt learning and multi-granularity aesthetic perception that capture more domain-agnostic principles, highlighting the practical advantage of our zero-shot approach for real-world applications.

Intra-dataset Setting. We further compare AesPrompt with

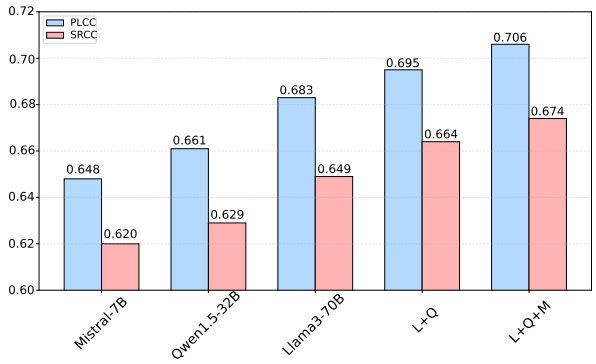
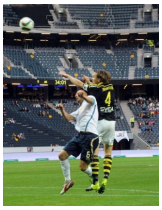


Fig. 6. Performance comparison of different LLM configurations for pseudo aesthetic distribution generation on AVA training set. Single models show increasing performance with model size, while the ensemble strategy (L+Q+M) achieves the best results. L: Llama3-70B, Q: Qwen1.5-32B, M: Mistral-7B.

Baseline	I don't get the connection to the challenge
Ours	The image lacks a clear focal point, resulting in a cluttered and unfocused composition the washed - out colors and washed - out quality further detract from the overall aesthetic appeal



Baseline	Awwwwwwww…
Ours	The image captures a dynamic moment of athleticism, with the players' expressions conveying a palpable sense of energy and athleticism, the composition could be improved by cropping the left side

Fig. 7. Qualitative comparison of aesthetic critique generation results. The baseline model, trained on raw noisy comments, produces uninformative or irrelevant comments. In contrast, our ACGM, trained with LLM-refined data, generates more comprehensive and interpretable aesthetic critiques that analyze specific visual elements and provide constructive feedback.

supervised IAA methods under the intra-dataset setting, where supervised methods are trained and tested on the same dataset while our method maintains its zero-shot setting. The performance comparison on AVA and TAD66K datasets is presented in Table IV, and the results on APDD dataset are shown in Table V. On the AVA database, our zero-shot approach achieves the second-highest performance, surpassed only by the recently proposed AesMamba [11]. For the TAD66K database, AesPrompt exhibits comparable results to several supervised methods. Most notably, on the artistic IAA dataset APDD [58], our method achieves exceptional performance (SRCC: 0.582) without any prior exposure to artistic images, closely rivaling state-of-the-art supervised artistic IAA methods [58], [62]. These results demonstrate that our zero-shot approach not only eliminates the need for aesthetic annotations but also achieves competitive or superior performance compared to supervised methods across diverse image domains.

D. Quality of Generated Pseudo Aesthetic Distribution

To systematically evaluate the quality of our generated pseudo-aesthetic distributions, we conduct a comprehensive analysis using the AVA training set as ground truth. Specifically, we convert the generated distributions into scalar scores using Equation (12) and measure their correlation with human annotations through PLCC and SRCC metrics, as shown in Fig. 6. Our analysis reveals two key findings. First, the capacity of LLMs shows a strong positive correlation with the quality of generated pseudo distributions. Among single models, Llama3-70B achieves the best performance, followed by Qwen1.5-32B and Mistral-7B. This trend suggests that larger models can better understand and assess aesthetic qualities from textual descriptions. More importantly, our proposed multi-LLM ensemble strategy significantly enhances the robustness of pseudo labels. The combination of all three models (L+Q+M) achieves the highest correlation scores, demonstrating that ensemble learning effectively leverages the complementary strengths of different LLMs for more reliable aesthetics assessment.

E. Quality of Generated Aesthetic Critiques

As discussed in Section III.C, we introduce a LLM-guided refinement strategy that transforms these raw comments into structured, informative aesthetic descriptions. The LLM processes raw comments through carefully designed prompts:

***#System**: You are an expert in the aesthetic critique of images.*

***#User**: You are an expert at image aesthetic critique, and you are looking at a picture. Your aesthetic description should be based on the following comments. Please evaluate the aesthetic quality of the image. Make sure to limit your comment to 50 words or less, the aesthetic comments are: [Comments].*

To validate the effectiveness of our LLM-guided refinement strategy in ACGM, we present qualitative examples of generated aesthetic critiques in Fig. 7. The example in Figure 7 demonstrates a critical issue when working with raw user comments: the baseline model sometimes produces irrelevant or contextually limited responses (e.g., "I don't get the connection to the challenge"). This is not a systematic failure but rather reflects the nature of the original DPChallenge platform, where users often commented in the context of specific photographic challenges or contests. Such platform-specific references can lead to generated critiques that lack aesthetic relevance. Our LLM-guided refinement strategy specifically addresses this issue by filtering out such contextual noise and focusing on transferable aesthetic insights, resulting in more consistent, relevant, and interpretable aesthetic critiques across diverse images. The enhanced quality of these critiques serves two important purposes in our framework: First, it demonstrates that our LLM-guided refinement strategy effectively transforms noisy user comments into structured, professional aesthetic evaluations. Second, these improved critiques provide richer semantic information for our aesthetic prompt learning process, contributing to the overall effectiveness of our AesPrompt framework.

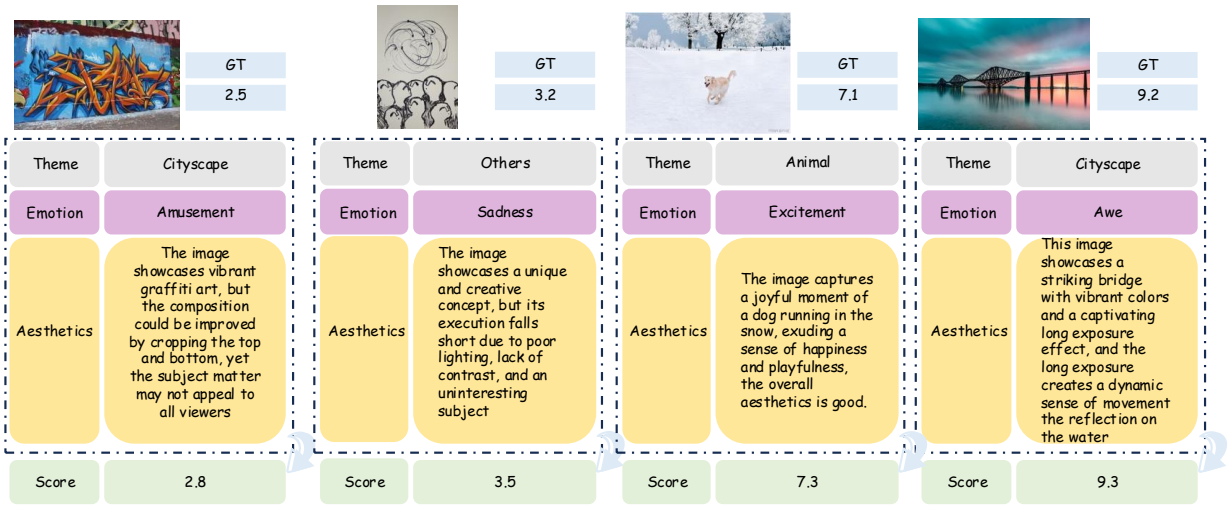


Fig. 8. Visualization of AesPrompt's human-like aesthetic assessment process. Following the cognitive patterns of human aesthetic perception, our model first identifies themes, then recognizes emotional responses, and finally provides reasoned aesthetic critiques, offering unprecedented interpretability in its decision-making process.

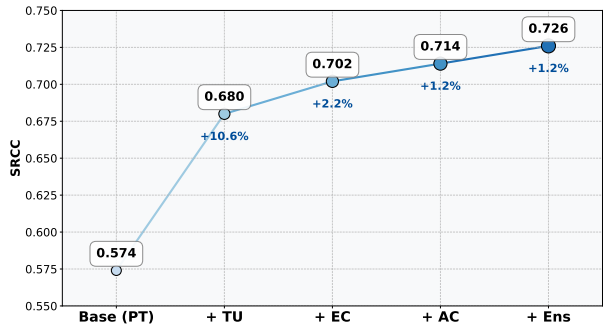


Fig. 9. Progressive improvement in SRCC performance with the addition of each component.

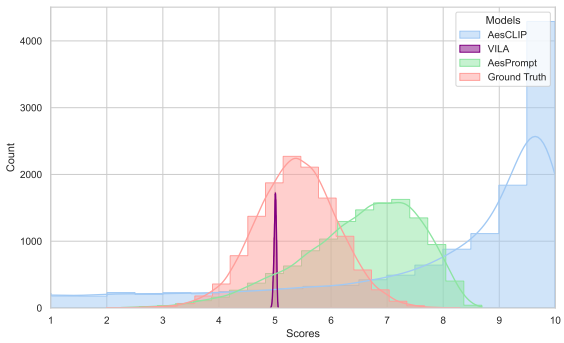


Fig. 10. Comparative analysis of aesthetic score distributions on the AVA test set. The plot demonstrates the predicted score distributions of different ZIAA models (AesCLIP [9], VILA [17], and our AesPrompt) against the ground truth distribution.

F. Ablation Study

To systematically investigate the effectiveness of each proposed component, we conduct comprehensive ablation studies on the AVA database, as illustrated in Fig. 9. We begin with a baseline model (PT) that uses CLIP pre-trained on AVA-comments database, and progressively incorporate different components while tracking the SRCC metric. The results reveal a clear pattern of consistent improvement. Starting from the baseline pre-trained model with an SRCC of 0.574, the introduction of theme understanding (TU) model brings a substantial improvement of 10.6%, reaching 0.680. This significant jump demonstrates the crucial role of theme understanding in aesthetics assessment. The addition of emotion classification (EC) model further enhances the SRCC to 0.702, suggesting that emotional perception provides complementary information for aesthetic judgment. When aesthetic critiques (AC) are integrated, the model achieves an SRCC of 0.714, validating our multi-granularity prompt learning strategy. Finally, the ensemble approach, which integrates prompts from diverse aesthetic attributes, marking a total improvement of 15.2% over the baseline. This steady progression in performance demonstrates that each component contributes meaningfully to the model's ability to assess image aesthetics, with the full model achieving the best results.

G. Visual Analysis

Our experimental results are visualized from two perspectives: the human-like aesthetic perception process and the score distribution analysis. Fig. 8 demonstrates how AesPrompt emulates human aesthetic perception through a hierarchical cognitive process, from basic theme recognition to emotional response, and finally to detailed aesthetic reasoning. AesPrompt not only provides interpretable insights into the model's decision-making process but also aligns with psychological studies on how humans evaluate visual aesthetics.

Fig. 10 presents a comparative analysis of score distributions generated by different ZIAA models on the AVA test set. The distribution analysis reveals several key findings: (1) VILA [17] shows a tendency to generate centralized predictions around score 5, indicating limited discriminative capability; (2) AesCLIP [9] demonstrates a notable bias towards high

Fig. 11. Examples of systematic failure patterns in AesPrompt.

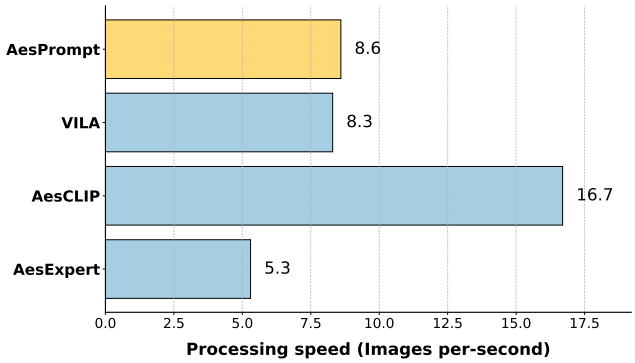


Fig. 12. Comparison of processing speeds (images per second) among different ZIAA methods.

scores, with predictions predominantly exceeding 8; (3) AesPrompt's predicted distribution exhibits the highest similarity to the ground truth, suggesting superior accuracy in aesthetic assessment.

While AesPrompt demonstrates strong generalization capabilities overall, our systematic analysis of failure cases revealed several consistent patterns where the model significantly underperforms. As shown in Fig. 11, we identified two primary categories of challenging images where our approach shows substantial limitations: 1) **Artistic works with historical context**: Classical paintings and traditional art forms (such as the still life example in Fig. 11-top) are consistently underrated by our model. The aesthetic value of these works often depends on art-historical knowledge, cultural significance, and technical traditions that are not well-represented in our training distribution of predominantly contemporary photographs. 2) **Images with deliberate technical deviations**: Works that intentionally employ what would be considered technical "flaws" in photography—such as extreme darkness, selective focus, or unconventional composition—as artistic devices (exemplified by Fig. 11-bottom) are frequently misinterpreted by our model as low-quality images rather than creative artistic choices. These patterns reveal a fundamental challenge in cross-domain aesthetic assessment: our multi-granularity prompts, while effective for natural images, struggle to bridge the significant domain gap between photographic conventions and artistic intentions. The model's aesthetic judgments remain anchored in photographic norms despite our efforts to incorporate broader aesthetic principles. This analysis suggests that future work should focus on developing domain-adaptive aesthetic prompts that can recognize and properly evaluate intentional deviations from conventional aesthetics across different visual domains.

H. Computational Efficiency

In addition to model performance and interpretability, computational efficiency is crucial for practical IAA applications. We analyzed the processing speed of our AesPrompt compared with state-of-the-art ZIAA methods [9], [17], [19] on the AVA database, with all tests conducted on an Intel Core i9-13900K CPU and NVIDIA GeForce RTX 4090 GPU. For fair comparison, we maintained consistent image resolutions across most methods, with VILA [17], AesCLIP [9], and our AesPrompt all using 224×224 input images, while AesExpert [19] requires a larger input size of 336×336. As shown in Fig. 12, AesCLIP achieves the highest efficiency, while our AesPrompt processes 8.6 images per second, comparable to VILA and significantly faster than AesExpert. The relatively lower speed of AesExpert can be attributed to its large-scale model components, while our approach achieves a balance between efficiency and functionality. Although AesPrompt incurs some computational overhead compared to AesCLIP due to its multi-granularity prompt learning strategy, this modest reduction in processing speed represents a reasonable trade-off given the substantial improvements in assessment accuracy and interpretability offered by our method, making it suitable for both offline evaluation tasks and interactive applications.

I. Limitations

While AesPrompt demonstrates strong performance across various image domains, there are several limitations worth noting. First, despite showing significant improvements over existing ZIAA methods, our approach still exhibits limited performance on AI-generated images. This performance gap highlights the substantial difference between natural image aesthetics and the unique visual characteristics of AI-generated content, suggesting that specific considerations for synthetic imagery may be needed. Second, the current implementation focuses on three aesthetic levels (Good, Average, Bad) for pseudo distribution generation, which may limit the granularity of aesthetic assessment. Expanding to more fine-grained aesthetic distributions could potentially improve performance but presents challenges in reliable label generation. Finally, while our framework incorporates aesthetic critiques generation for enhanced interpretability, there remains room for improvement in providing more specific and actionable feedback for aesthetic enhancement. Future research could explore methods to generate more targeted suggestions for improving the aesthetic quality of images across different domains.

V. Conclusion

This paper presents AesPrompt, a novel framework for Zero-Shot Image Aesthetics Assessment (ZIAA) that bridges the gap between human aesthetic perception and computational assessment. Our investigation has revealed several key

findings. First, the success of multi-granularity prompt learning validates that effective aesthetic assessment requires hierarchical understanding from theme to emotion to aesthetics, suggesting future ZIAA methods should consider the cognitive process of human perception. Second, the effectiveness of our multi-LLM ensemble in generating reliable pseudo aesthetic distributions demonstrates the potential of leveraging linguistic knowledge for visual assessment tasks. Third, AesPrompt's strong performance across natural, artistic, and AI-generated images indicates that well-designed prompt learning strategies can enable effective aesthetic knowledge transfer across diverse visual domains. We believe that AesPrompt would inspire the further advancements in ZIAA and serve as a catalyst for the future research of low-level vision MLLM for better zero-shot assessment ability.

REFERENCES

[1] L. Zhang, M. Xu, J. Yin, C. Zhang, and L. Shao, "Weakly supervised complets ranking for deep image quality modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5041–5054, 2020.

[2] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang *et al.*, "Rich human feedback for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 401–19 411.

[3] T. Zhou, Z. Cai, F. Liu, and J. Su, "In pursuit of beauty: Aesthetic-aware and context-adaptive photo selection in crowdsensing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9364–9377, 2023.

[4] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 652–36 663, 2023.

[5] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[6] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[7] H. Chen, F. Shao, W. Jing, H. Wang, and Q. Jiang, "Cross-modal hierarchical knowledge distillation for image aesthetics assessment," *IEEE Transactions on Multimedia*, pp. 1–14, 2024.

[8] S. He, A. Ming, S. Zheng, H. Zhong, and H. Ma, "Eat: An enhancer for aesthetics-oriented transformers," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1023–1032.

[9] X. Sheng, L. Li, P. Chen, J. Wu, W. Dong, Y. Yang, L. Xu, Y. Li, and G. Shi, "Aesclip: Multi-attribute contrastive learning for image aesthetics assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1117–1126.

[10] F. Guan, X. Li, Z. Yu, Y. Lu, and Z. Chen, "Q-mamba: On first exploration of vision mamba for image quality assessment," *arXiv preprint arXiv:2406.09546*, 2024.

[11] F. Gao, Y. Lin, J. Shi, M. Qiao, and N. Wang, "Aesmamba: Universal image aesthetic assessment with state space models," in *ACM Multimedia 2024*, 2024.

[12] Y. Shu, Q. Li, L. Liu, and G. Xu, "Semi-supervised adversarial learning for attribute-aware photo aesthetic assessment," *IEEE Transactions on Multimedia*, vol. 26, pp. 4086–4096, 2024.

[13] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, "Deep active learning with contaminated tags for image aesthetics assessment," *IEEE Transactions on Image Processing*, 2018.

[14] Y. Li, Y. Yang, H. Li, H. Chen, L. Xu, L. Li, Y. Li, and Y. Guo, "Transductive aesthetic preference propagation for personalized image aesthetics assessment," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 896–904.

[15] Z. Yang, L. Li, Y. Yang, Y. Li, and W. Lin, "Multi-level transitional contrast learning for personalized image aesthetics assessment," *IEEE Transactions on Multimedia*, 2023.

[16] J. Hou, W. Lin, G. Yue, W. Liu, and B. Zhao, "Interaction-matrix based personalized image aesthetics assessment," *IEEE Transactions on Multimedia*, vol. 25, pp. 5263–5278, 2023.

[17] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang, "Vila: Learning image aesthetics from user comments with vision-language pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 041–10 051.

[18] G. Wang, Y. Tan, H. Lin, and C. Zhang, "Keep knowledge in perception: Zero-shot image aesthetic assessment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 8311–8315.

[19] Y. Huang, X. Sheng, Z. Yang, Q. Yuan, Z. Duan, P. Chen, L. Li, W. Lin, and G. Shi, "Aesexpert: Towards multi-modality foundation model for image aesthetics perception," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5911–5920.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[22] G. Cabbai, C. Kühnapfel, J. Fingerhut, L. Kaltwasser, J. J. Prinz, and M. Pelowski, "Emotion, embodiment, and aesthetic appraisal: The impact of interoceptive abilities and art type," *Psychology of Aesthetics, Creativity, and the Arts*, 2023.

[23] G. Valenzise, C. Kang, and F. Dufaux, "Advances and challenges in computational image aesthetics," *Human perception of visual information: Psychological and computational perspectives*, pp. 133–181, 2022.

[24] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proceedings of the European Conference on Computer Vision*. Springer, 2006, pp. 288–301.

[25] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, pp. 419–426.

[26] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 386–399.

[27] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver, "The role of image composition in image aesthetics," in *Proceedings of the IEEE International Conference on Image Processing*, 2010, pp. 3185–3188.

[28] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2206–2213.

[29] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 457–466.

[30] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[31] S. Ma, J. Liu, and C. Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 4535–4544.

[32] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 497–506.

[33] Y. Shu, Q. Li, L. Liu, and G. Xu, "Privileged multi-task learning for attribute-aware aesthetic assessment," *Pattern Recognition*, vol. 132, p. 108921, 2022.

[34] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 3898–3910, 2020.

[35] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 8475–8484.

[36] L. Li, Y. Huang, J. Wu, Y. Yang, Y. Li, Y. Guo, and G. Shi, "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

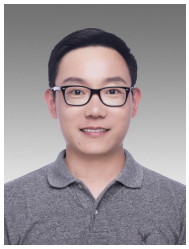
[37] L. Li, X. Sheng, P. Chen, J. Wu, and W. Dong, "Towards explainable image aesthetics assessment with attribute-oriented critiques generation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[38] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang, "Joint image and text representation for aesthetics analysis," in *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 2016, pp. 262–266.

[39] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified

image aesthetic prediction tasks," *IEEE Transactions on Multimedia*, vol. 23, pp. 611–623, 2020.
[40] Y. Niu, S. Chen, B. Song, Z. Chen, and W. Liu, "Comment-guided semantics-aware image aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1487–1492, 2023.
[41] L. Li, T. Zhu, P. Chen, Y. Yang, Y. Li, and W. Lin, "Image aesthetics assessment with attribute-assisted multimodal memory network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.
[42] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2023.
[43] Y. Huang, Q. Yuan, X. Sheng, Z. Yang, H. Wu, P. Chen, Y. Yang, L. Li, and W. Lin, "AesBench: An expert benchmark for multimodal large language models on image aesthetics perception," *arXiv preprint arXiv:2401.08276*, 2024.
[44] Z. Zhou, Q. Wang, B. Lin, Y. Su, R. Chen, X. Tao, A. Zheng, L. Yuan, P. Wan, and D. Zhang, "Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark," 2024.
[45] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
[46] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.
[47] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.
[48] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[49] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
[50] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
[51] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677–3686.
[52] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 383–20 394.
[53] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.
[54] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou, and X. Zhou, "Aesthetic attributes assessment of images," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, p. 311–319.
[55] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 662–679.
[56] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.
[57] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *Proceeding of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 942–948.
[58] X. Jin, Q. Qiao, Y. Lu, S. Gao, H. Huang, and G. Li, "Paintings and drawings aesthetics assessment with rich attributes for various artistic categories," in *International Joint Conferences on Artificial Intelligence*, 2024.
[59] J. D. Pressman, K. Crowson, and S. C. Contributors, "Simulacra aesthetic captions," Stability AI, Tech. Rep. Version 1.0, 2022, url https://github.com/JD-P/simulacra-aesthetic-captions .
[60] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2019.
[61] H. Zhu, Y. Zhou, L. Li, Y. Li, and Y. Guo, "Learning personalized image aesthetics from subjective and objective attributes," *IEEE Transactions on Multimedia*, vol. 25, pp. 179–190, 2021.
[62] R. Yi, H. Tian, Z. Gu, Y.-K. Lai, and P. L. Rosin, "Towards artistic image aesthetics assessment: a large-scale dataset and a new method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 388–22 397.

Xiangfei Sheng (Student Member, IEEE) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2021. He is currently pursuing a Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His current research interests include multimedia quality assessment and processing, visual perception modeling.



Leida Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Full Professor with the School of Artificial Intelligence, Xidian University, China. His research interests include multimedia quality assessment, computational aesthetics and visual sentiment analysis. He served as Area Chair for ACM Multimedia 2025, SPC for IJCAI 2019-2021 and 2025. He is currently an Associate Editor of IEEE Transactions on Image Processing, Journal of Visual Communication and Image Representation and EURASIP Journal on Image and Video Processing.

Pengfei Chen received the B.S. degree from Xidian University, Xi'an, China, in 2014. He received Ph.D. from the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China in 2023. He is currently a faculty member with the School of Artificial Intelligence, Xidian University. His research interests include image/video quality assessment, video quality of experience, and domain adaptation/generalization.

Li Cai received his BS degree from China University of Mining and Technology in 2004 and his master's degree from Chongqing University in 2010. From 2014 to 2015, he was a visiting scholar at the School of Electrical Engineering, Southeast University. He is currently a professor and the head of the Department of Electrical Engineering at Chongqing Three Gorges University. His research interests include key technologies of electric vehicle battery packs and intelligent driving technologies. He currently serves as a young editorial board member of the Battery Journal and the director of the Wanzhou District Technology Innovation Center.

Giuseppe Valenzise (Senior Member, IEEE) received the Ph.D. degree in information technology with the Politecnico di Milano, Torino, Italy, in 2011. He is currently a CNRS Researcher with Laboratoire des Signaux et Systèmes (L2S), Université Paris-Saclay, CentraleSupélec, Gir-sur-Yvette, France, where he is the Head of the Multimedia and Networking Team. In 2012, he joined the French Centre National de la Recherche Scientifique (CNRS) as a permanent Researcher, first with the Laboratoire Traitement et Communication de l'Information (LTCI) Telecom Paristech, and in 2016 with L2S. He is the coauthor of more than 100 research publications and of several award-winning papers. His research interests include different fields of image and video processing, traditional and learning-based image and video compression, immersive video (light fields, point clouds), image/video quality assessment, high dynamic range imaging, and applications of machine learning to image and video analysis. He was the recipient of the EURASIP Early Career Award in 2018 for "significant contributions to video coding and analysis". He was/is an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON IMAGE PROCESSING, Elsevier Signal Processing: Image communication. He is the Chair of the MMSP Technical Committee of the IEEE Signal Processing Society for the term 2024–2025, and he was a member of the Technical Area Committee on Visual Information Processing of EURASIP from 2018 to 2023.