Open World Image Aesthetic Assessment

Mingxiang Liao[1]* Tianren Ma[2]* Xijin Zhang[1]
[1]ByteDance Inc. [2]University of Chinese Academy of Sciences
{liaomingxiang, zhangxijin}@bytedance.com matianren18@mails.ucas.ac.cn

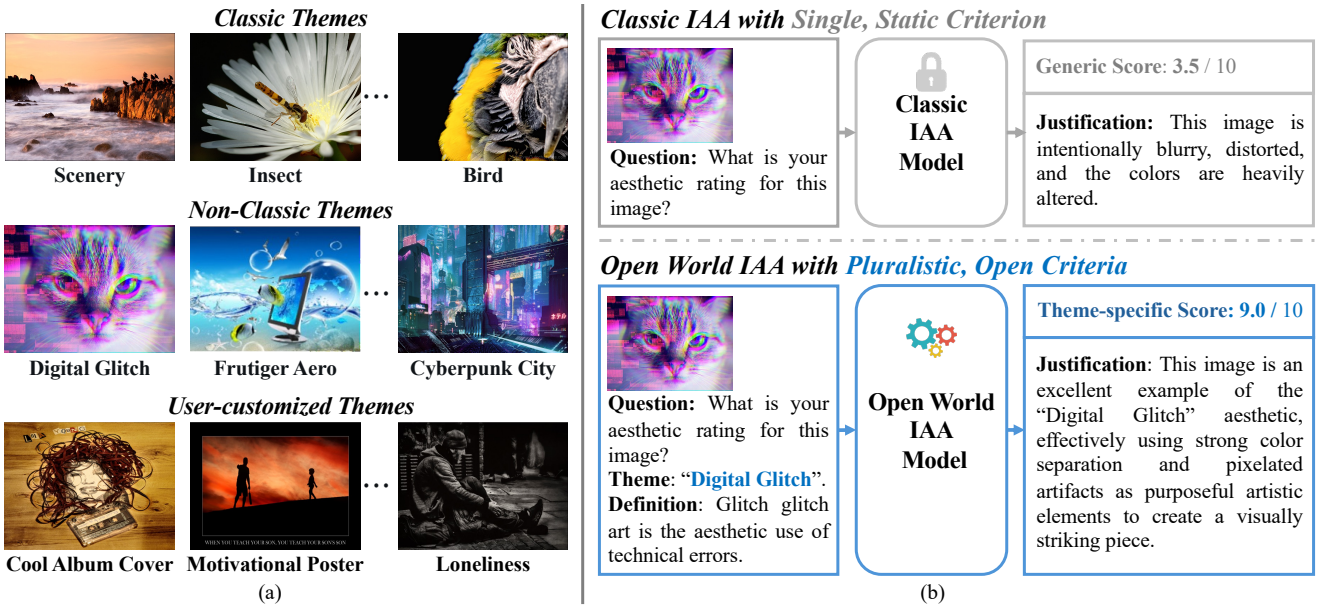

Figure 1. **Comparison between Classic Image Aesthetic Assessment (IAA) and Open-World IAA.** (a) Examples of diverse aesthetic themes including classic themes, non-classic themes, and user-customized themes. (b) Classic IAA applies a single, static criterion, yielding a low generic score that fails to capture theme-specific aesthetic values. Our Open-World IAA evaluates images with open, theme-specific criteria, providing justified assessments that accurately reflect aesthetic quality for the given theme (9.0/10 for "Digital Glitch").

Abstract

Prevailing Image Aesthetic Assessment (IAA) models assume a single, generic aesthetic criterion, yet real-world aesthetics are fundamentally pluralistic (diverse themes, unique criteria) and open (new themes constantly emerging). To bridge this gap, we propose the Open-World IAA (OW-IAA) task, requiring models to exhibit both versatility across diverse seen themes and generalization to unseen ones. To address this, we propose Induce-and-Adapt, a novel framework that first induces a generalizable reasoning policy via jointly optimizing for score prediction and reasoning and then efficiently adapts to new themes by optimizing policy against crowd-simulated preferences—achieved without human annotation. To evaluate this task, we introduce OA-Bench, the first benchmark for OW-IAA, which reveals current methods are ill-equipped for this challenging setting. Our method builds new state-of-the-art on both OA-Bench and TAD66K, and achieves 0.102 PLCC improvement over prior best. It also demonstrates a synergistic lift by improving generalization to unseen themes without sacrificing versatility on seen ones. Project page: github.com/MingXiangL/OW-IAA.

1. Introduction

Driven by the advancements in deep learning, the field of Image Aesthetic Assessment (IAA) has made significant strides. However, prevailing methods are largely predicated on a core assumption: the existence of a singular, generic aesthetic criterion [12, 25, 38]. This assumption overlooks two fundamental challenges inherent to IAA. First, **aesthetic pluralism**, where distinct aesthetic

*Equal Contribution

themes like "digital glitch art" possess their own criteria that can diverge from or even contradict a generic criterion (Fig. 1(b)). Second, **openness**, the perpetual emergence of new themes—from novel styles like "cyberpunk city" to user-customized themes—each introducing its own unique aesthetic criteria (Fig. 1(a)).

To overcome this limitation, we propose a more realistic and challenging task: Open World Image Aesthetic Assessment (OW-IAA). This task aims to shift models away from fitting a single, static aesthetic criterion towards cultivating genuine aesthetic understanding and reasoning capabilities. An ideal OW-IAA model must meet a dual mandate: (1) **Versatility**: Master the distinct aesthetic criteria across a large and diverse set of hundreds of training themes. (2) **Generalization**: Accurately evaluate themes unseen during training, based solely on their textual descriptions.

Existing paradigms like Personalized (PIAA) and Theme-aware (TIAA) IAA are ill-equipped for the open-world challenge: PIAA often overlooks theme-specific criteria in favor of user taste, while TIAA is inherently confined to a closed set of themes.

To overcome these challenges, we propose the "Induce-and-Adapt" framework (IAF), a two-stage approach for OW-IAA. **Induction stage**: We train a foundational policy on seen themes, jointly optimizing for score prediction and Chain-of-Thought (CoT) reasoning to instill generalizable aesthetic reasoning. **Adaptation stage**: At test time, when faced with unseen themes, the model performs unsupervised adaptation. Leveraging the generalizable reasoning ability from the induction stage, the model first simulates crowd aesthetic preferences by aggregating multiple rollouts. This informed exploration generates robust pseudo-labels, which then serve as reward signals to efficiently refine its policy online using the GRPO algorithm [5, 29, 50], thereby enhancing the generalization of the policy.

To enable reliable evaluation of a model's versatility and generalization, we constructed OA-Bench, the first benchmark for OW-IAA. We began by restructuring the AVA dataset, curating 871 distinct aesthetic themes from its 1,477 "challenges". To systematically assess performance across these themes, our design incorporates two key components: (1) **The Aesthetic Criterion Deviation (ACD) metric:** We introduce ACD to quantify the deviation of each theme's aesthetic criterion from a generic baseline (the global AVA average). This allows us to analyze how performance correlates with thematic distinctiveness. (2) **A stratified split protocol:** Based on ACD, we partition themes into high, medium, and low-deviation tiers. From these tiers, we sample a large seen set (785 themes) to test versatility and a challenging unseen set (87 themes) to test generalization. This design enables a comprehensive evaluation of a model's core open-world capabilities. Our benchmarks show that while SOTA models exhibit some versatility on the seen set, their performance degrades substantially on the unseen and high-deviation themes, confirming the unique challenge posed by OW-IAA.

Extensive experiments validate our framework's effectiveness. IAF achieves new SOTA results on both OA-Bench and TAD66K [6], with a +0.102 Pearson correlation gain over prior best methods [12, 38] on the challenging Unseen High-Deviation set. Crucially, we observe a **synergistic lift**: adapting to unseen themes improves generalization on novel themes while simultaneously enhancing performance on seen themes. It shows that the adaptation stage refines the generalizable reasoning policy rather than creating zero-sum trade-offs. This adaptation is highly data-efficient, requiring only one sample per unseen theme.

The main contributions of this paper are as follows:

- We propose the Open-World Image Aesthetic Assessment task to capture the pluralistic and open nature of real-world aesthetics.
- We introduce IAF, a novel framework that first induces a generalizable aesthetic policy by joint score prediction and reasoning, then adapts to unseen themes via efficient unsupervised refinement guided by crowd simulation.
- We present OA-Bench, the first benchmark to systematically evaluate both versatility on seen themes and generalization to unseen ones.
- IAF achieves SOTA on both OA-Bench and TAD66K, and shows a synergistic lift where adaptation to unseen themes enhances performance on seen ones.

2. Related Works

2.1. Classic Image Aesthetic Assessment

Classic Image Aesthetic Assessment (CIAA) operates on the assumption of a universal aesthetic criterion [9, 15–18, 25, 27, 36, 37, 45]. This field has evolved from early benchmarks like the AVA dataset [25] to recent instruction-tuning datasets for MLLMs, such as AesExpert [8] and data from photography experts [27]. Methodologies are primarily divided into regression-based approaches that predict scores (e.g., HLA-GCN [30], AesMamba [3], NIMA [32]) and MLLM-based methods that enhance flexibility and interpretability (e.g., Q-Align [38], Q-Instruct [37], Compare2Score [49]). However, the foundational "universal criterion" assumption prevents CIAA from addressing the pluralistic and often contradictory aesthetic criteria of the real world. OW-IAA posits that aesthetic criteria are intrinsically pluralistic and open, underscoring the necessity of versatility on training themes and generalization on unseen themes.

2.2. Personalized Image Aesthetic Assessment

Personalized Image Aesthetic Assessment (PIAA) posits that aesthetic criteria are user-dependent, aiming to learn a

dedicated model for each individual. Datasets have evolved from user preference annotations (FLICKR-AES [28]) to include personality traits (PARA [42]) and, more recently, specific aesthetic themes like artwork (LAPIS [23]) and physique images (PhysiqueAA50K [48]). Technical approaches include collecting user feedback (Usar [21]), incorporating personality traits [10], and enhancing learning efficiency through methods like task vector customization [44] or transductive preference propagation (TAPPPIA[14]). The key difference from OW-IAA is that PIAA attributes aesthetic differences to the user ("Do you like this image?"), whereas OW-IAA attributes them to the theme ("Is this image good for this theme?").

2.3. Theme-aware Image Aesthetic Assessment

Theme-aware Image Aesthetic Assessment (TIAA) is founded on the principle that theme significantly impacts aesthetics, proposing that an image's theme should be identified before evaluation [20]. To facilitate this, researchers constructed large-scale, multi-theme datasets like TAD66K [6], which contains 47 popular themes. This led to the development of models such as TANet [6], designed to adaptively extract theme information and establish specific perception rules. Other diverse strategies have also been explored; for instance, TAVAR [11] simulates theme-aware judgment through "two-level reasoning," while Jia et al.[2] fuse aspect ratio with theme information to address the feature loss problem. This focus extended to related concepts like "scenes," (e.g. SPAQ dataset[2]), and "styles," where style-specific features are fused with general aesthetic features for artistic images [43]. Despite this progress, TIAA's basic limitation is its confinement to the predefined set of "seen" themes from the training phase. In contrast, OW-IAA targets an "open world" setting, demanding generalization to "unseen" themes.

3. Problem Definition

Prevailing IAA methods assume a singular, static aesthetic criterion, which is inadequate for the pluralistic and open nature of real-world aesthetics, where different themes have distinct criteria and new themes continually emerging.

To address this, we propose the Open-World Image Aesthetic Assessment (OW-IAA) task, formally defined as:
Given a set of seen themes $\mathcal{T}_{\text{seen}} = \{t_1, ..., t_n\}$, where each theme $t \in \mathcal{T}_{\text{seen}}$ contains labeled image-score pairs (x_i, s_i), and theme descriptors (names and textual descriptions). At test time, the model encounters a disjoint set of unseen themes $\mathcal{T}_{\text{unseen}}$, where $\mathcal{T}_{\text{seen}} \cap \mathcal{T}_{\text{unseen}} = \emptyset$. The goal is to learn a model that: (1) accurately scores images from $\mathcal{T}_{\text{seen}}$, and (2) generalizes to images from $\mathcal{T}_{\text{unseen}}$ using only their theme descriptors and unlabeled images.

This requires the model to possess two core capabilities: (1) **Versatility** to master the diverse aesthetic criteria across seen themes. (2) **Generalization** to understand and adapt to the aesthetic criteria of new themes at test time.

Algorithm 1 Pairwise CoT Construction

Require: Dataset $\mathcal{D} = \{(x_i, s_i, t_i, d_i)\}_{i=1}^{M}$, where x_i is an image, s_i is its score, t_i is the theme name, d_i is the theme description, and $\mathcal{D}_{t_i}$ is the subset of $\mathcal{D}$ for theme t_i.
Ensure: CoT dataset $\mathcal{C} = \{(x_i, s_i, t_i, d_i, \text{cot}_i)\}_{i=1}^{M}$

CoT_map $\leftarrow \{\}$ ▷ A map to store validated CoTs
for $i \leftarrow 1$ to M **do**
  Let $sample_i = (x_i, s_i, t_i, d_i)$
  while $x_i \notin$ CoT_map **do**
    $sample_j \leftarrow$ random_sample$(\mathcal{D}_{t_i} \setminus \{sample_i\})$
    $(x_j, s_j, t_j, d_j) \leftarrow sample_j$
    ▷ Call GPT-4o to analyze and order x_i, x_j
    $(\text{cot}_i, \text{cot}_j, \text{pred\_order}) \leftarrow \text{GPT}(x_i, x_j, t_i, d_i)$
    gt_order $\leftarrow$ get_order(s_i, s_j)
    if pred_order == gt_order **then**
      CoT_map$[x_i] \leftarrow \text{cot}_i$
      CoT_map$[x_j] \leftarrow \text{cot}_j$
    end if
  end while
end for
$\mathcal{C} \leftarrow \{(x_i, s_i, t_i, d_i, \text{CoT\_map}[x_i]) \mid x_i \in \mathcal{D}\}$
return $\mathcal{C}$

4. Induce-and-Adapt Framework

To fulfill the requirements of versatility and generalization for OW-IAA, we propose the Induce-and-Adapt framework (IAF). Transcending static criterion fitting, our framework cultivates generalizable aesthetic reasoning capability through an extensible learning process in two stages (Fig. 2): (1) **Aesthetic Criterion Induction** to build a foundational policy that, through induction on training themes, masters their diverse criteria (Versatility) and establishes the generalizable reasoning priors for adaptation (Generalization-base), and (2) **Test-time Aesthetic Policy Adaptation** to adapt this policy to unseen themes.

4.1. Aesthetic Criterion Induction (ACI)

The primary objective of ACI is twofold: (1) to master the diverse criteria of seen themes (Versatility), and (2) to induce a generalizable aesthetic reasoning capability required for adaptation (Generalization-base). Achieving this requires the model to not only learn what score to output, but also how to reason about aesthetics based on the provided criteria.

To achieve this, we incorporate a CoT-Guided prediction objective into our training. By prompting the model to generate an explicit textual rationale before its final prediction,

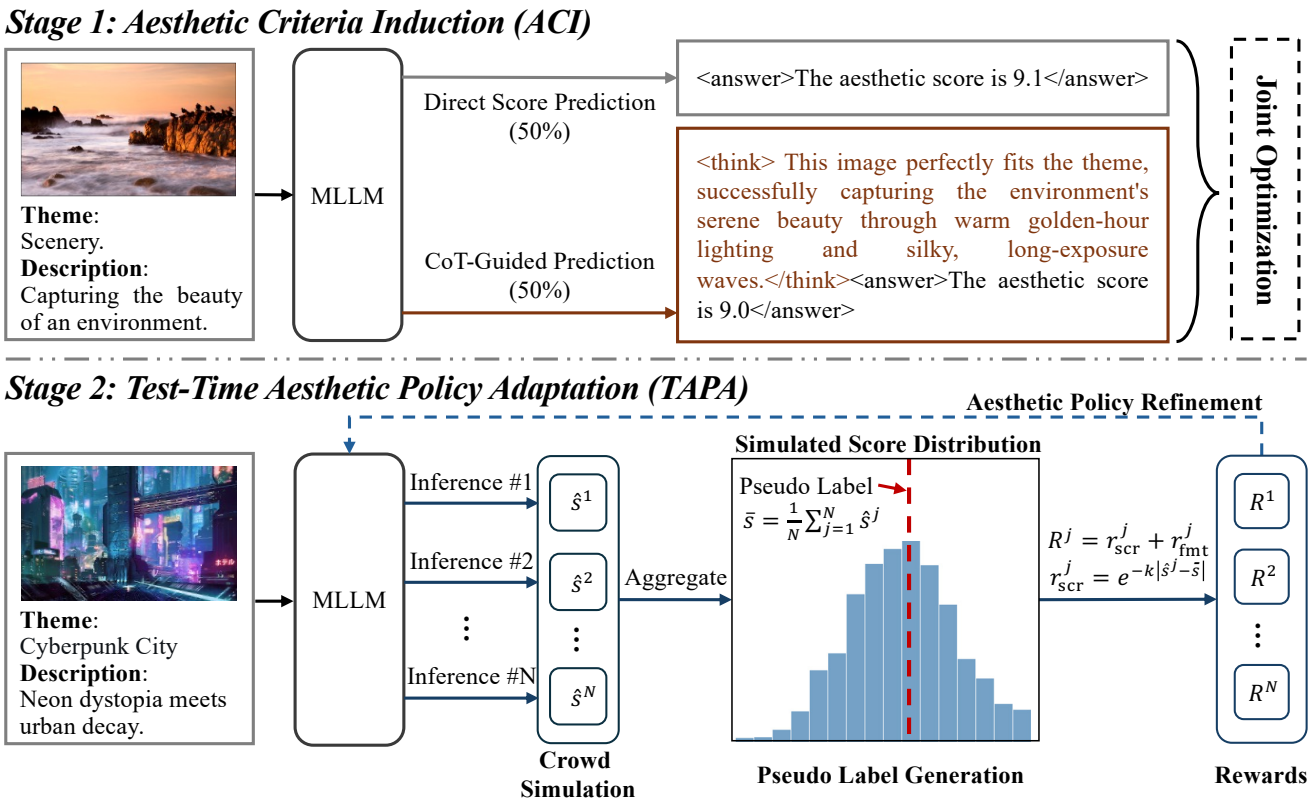

Figure 2. **The Induce-and-Adapt Framework. Stage 1 (Aesthetic Criteria Induction, ACI):** Learns a foundational MLLM policy from seen themes. **Stage 2 (Test-Time Aesthetic Policy Adaptation, TAPA):** Adapts this foundational policy to new aesthetic criteria for unseen themes guided by pseudo labels generated by crowd simulation.

we guide it to learn the analytical steps of aesthetic judgment. The effectiveness of this approach, however, hinges on the quality of the CoT data. The reasoning must be faithful to the ground-truth aesthetic scores.

To this end, we propose the **Pairwise CoT Construction** strategy (Algorithm 1). This approach involves sampling image pairs from the same theme and prompting GPT-4o to generate a comparative CoT analysis. Specifically, for each pair, GPT-4o is instructed to first produce a detailed, independent analysis of each image, evaluating theme-specific aesthetic aspects such as composition, lighting, and color. It then uses these individual analyses to formulate a final comparative judgment and quality ranking. We validate these annotations by checking if the resulting ranking aligns with the pair's ground-truth scores. Only the CoT from correctly ranked pairs is retained. This pairwise mechanism filters out ∼30% CoT analyses with misaligned conclusions, thus improving consistency with the ground-truth scores.

With the high-quality CoT annotations generated, we employ a **joint optimization strategy**, as depicted in Fig. 2. This strategy trains the model on two complementary tasks in a 1:1 ratio. The **Direct Score Prediction** task anchors the model's output to precise numerical values, ensuring scoring accuracy. Concurrently, the **CoT-Guided Prediction** task leverages our generated annotations to explicitly teach the model the "why" behind a judgment, requiring it to first produce a rationale before predicting the score.

Jointly optimizing these two tasks allows the model to achieve both ACI objectives: Direct Score Prediction ensures accuracy on seen themes (Versatility), while CoT-Guided Prediction equips the model with a generalizable aesthetic reasoning prior. This prior enables plausible zero-shot judgments on unseen themes and offers a robust foundational policy for the subsequent adaptation.

4.2. Test-Time Aesthetic Policy Adaptation (TAPA)

After acquiring generalizable aesthetic reasoning capabilities through ACI, we conduct test-time adaptation to further align the foundational policy with the aesthetic criteria of unseen themes. This refinement is crucial, as the foundational policy's zero-shot judgments, while robust, cannot fully capture the distinctive criterion of a novel theme. We propose a Test-time Aesthetic Policy Adaptation method comprising two components: Pseudo Label Generation and Aesthetic Policy Refinement, as shown in Fig. 2. The entire process requires no manual annotation.

Pseudo Labeling Generation. To obtain pseudo labels for new themes, we leverage the foundational policy to simulate crowd scores. We refer to this process as **Crowd Simulation**. Specifically, for each image in an unseen theme, we perform N independent inferences to obtain a set of scores $\{s^1, s^2, \ldots, s^N\}$. Sampling multiple times helps capture the uncertainty in aesthetic judgment and the diversity of human preferences. We then average these scores to derive the pseudo label $\bar{s} = \frac{1}{N}\sum_{j=1}^{N} s^j$, which serves as the supervision signal for subsequent policy refinement.

Aesthetic Policy Refinement. Based on the pseudo labels from crowd simulation, we refine the model's policy via reinforcement learning to adapt to new aesthetic preferences. The reward function combines score reward (r_{scr}) and format reward (r_{fmt}): $R = r_{\text{scr}} + r_{\text{fmt}}$.

Score Reward. To align the policy's predicted score $\hat{s}_i$ with pseudo label $\bar{s}$, we define:

$$r_{\text{scr}} = \exp(-k \cdot |\hat{s} - \bar{s}|) \tag{1}$$

Where hyperparameter k controls deviation tolerance: small errors receive mild penalties while large deviations are severely suppressed. Empirically, we set $k = 5$.

Format Reward. r_{fmt} ensures output format compliance.

Optimization. We employ GRPO [5, 29] to maximize expected reward. GRPO groups multiple samples from the same image and performs stable updates via group-relative advantage estimation. For efficiency, we reuse the N inferences from crowd simulation as group samples, avoiding redundant computation. Notably, benefiting from the quality of pseudo labels and the reward design, this method demonstrates **strong efficiency**, rapidly adapting to new themes with only 1 sample per theme and 10 iterations.

5. OA-Bench

To evaluate model capabilities in the OW-IAA setting, we propose OA-Bench. We first introduce a dataset with diverse themes curated from AVA (Sec. 5.1). We then propose Aesthetic Criterion Deviation (ACD), a novel metric to quantify the deviation between theme-specific criteria and the "generic" criterion (Sec. 5.2). Finally, we design an evaluation protocol (Sec. 5.3) that leverages ACD to partition themes into "Seen/Unseen" and "Low/Medium/High" deviation levels. This protocol enables a comprehensive assessment of both a model's versatility on seen themes and its generalization to unseen ones.

5.1. Theme Annotation

We construct the dataset by augmenting the widely-used AVA dataset with rich theme annotations. To achieve this, we first collected challenge descriptions from DPChallenge [1], obtaining descriptions for 1,477 challenges present in AVA. For precise thematic grouping, we employ the SOTA

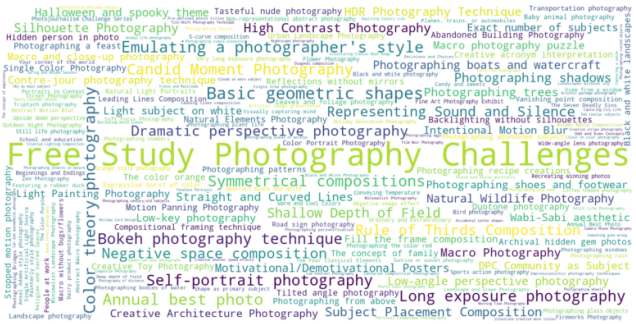

Figure 3. Word cloud of themes in OA-Bench. Font size indicates the image counts of corresponding theme.

text embedding model Qwen3-Embedding-8B [46] to extract the semantic embeddings of these descriptions. The resulting vectors were then clustered using HDBSCAN [24]. Subsequently, we leveraged Gemini-2.5 Pro [4] to generate a representative label and a theme description for each cluster, forming the initial set of theme labels. Finally, all theme labels were manually reviewed and verified to ensure accuracy. This process resulted in a final set of 871 unique themes. As shown in Fig. 3, the themes are diverse, including general aesthetic challenges ("Free Study Photography Challenge"), specific photographic techniques ("Rule of Thirds Composition"), subject-focused photography ("People at work"), and emotional expression ("Motivational/Demotivational Posters"). This thematic diversity ensures that the benchmark can comprehensively evaluate aesthetic models across a wide spectrum of classic themes, novel themes and user-customized themes.

5.2. Aesthetic Criterion Deviation

To quantify the deviation between theme-specific and "generic" aesthetic criteria, we first establish a generic aesthetic baseline. This is achieved by training a regression-basd classic images aesthetic assessment model (θ) on the aggregated data from all themes within the AVA dataset. The model consists of a two-layer MLP built upon the siglip2-base backbone [33]. This baseline model was trained by fine-tuning both the backbone and the MLP on the aggregated data from all themes, achieving a strong average Pearson correlation of 0.75 across the test sets of all themes, thus establishing it as a robust generic baseline. We then define the ACD for theme t as the baseline's performance drop on that theme's test set, calculated as:

$$\textbf{ACD}(t) = 1 - \textbf{PLCC}_t(\theta), \tag{2}$$

where $\textbf{PLCC}_t(\theta)$ represents the Pearson correlation performance of model θ on the test set of theme t. A higher ACD value indicates a greater deviation between the aesthetic criterion of theme t and the generic criterion.

Table 1. Dataset statistics of OA-Bench. This table shows theme counts by deviation level (High, Medium, Low) and the total number of images for the Train, Seen Test, and Unseen Test sets.

	Theme Counts				#Image
	Low	Medium	High	Total	
Train	157	471	157	785	205743
Seen Test	37	213	17	267	22387
Unseen Test	17	52	17	86	24234

5.3. Evaluation Protocol

To evaluate OW-IAA capabilities, we design an evaluation protocol by first partitioning image themes based on their visibility and criterion deviation (quantified by the ACD score, Sec. 5.2), and subsequently constructing the final datasets for evaluation(See Table. 1).

Theme Partitioning. First, we label **aesthetic deviation** of themes by ACD score following a standard partition by quintiles: the top 20%, middle 60%, and bottom 20% are labeled as High, Medium, and Low Level, respectively. Second, for **versatility**, we simulate the open-world challenge by performing a stratified sampling based on the deviation level. We select 10% of themes from each of the High, Medium, and Low deviation groups to form the *Unseen Themes* set, while the remaining 90% constitute the *Seen Themes set*.

Dataset Construction. The protocol consists of three final datasets as shown in Table 1. The *Seen Test Set* is composed of 10% of images from those themes in the Seen Themes set with over 500 images, resulting in a set of 22,387 images from 267 themes. The *Train Set* is defined as all remaining images from the Seen Themes set, totaling 205,743 images across 785 themes. Finally, the *Unseen Test Set* is composed of all images from the Unseen Themes set, containing 24,234 images across 86 themes; this set serves as the core testbed for generalization.

Evaluation Procedure. The model is trained on the Train Set. Its performance is first evaluated on the Seen Test Set to test its versatility on known themes. Subsequently, the Unseen Test Set is used to quantify the model's generalization to novel aesthetic criteria. For each theme in the test set, we test the correlation of each theme, and them report the mean correlation across themes.

6. Experiments

6.1. Experimental Setup

Benchmarks. Our experiments are primarily based on our *OA-Bench* (Sec. 5), our benchmark designed for the OW-IAA task to evaluate model versatility (seen themes) and generalization (unseen themes). It contains a train set (205,743 images / 785 themes), a seen test set for versatility (22,387 images / 267 themes), and an unseen test set for generalization (24,234 images / 87 themes). For analysis, all test sets are partitioned into "High," "Medium," and "Low" deviation levels via our ACD metric. We also validate our method on *TAD66K* [6], an established large-scale, theme-aware benchmark. We use it to test our model's performance against SOTA methods in this TIAA setting.

Baselines. We compare our IAF framework against several SOTA baselines. These include: (1) Vanilla MLLMs (e.g., GPT-4o [26], Qwen2-VL [34]); (2) Text-conditioned Reward Models (e.g., HPSv3 [22], ImageReward [40]); and (3) SOTA IAA Models (e.g., Q-align [38], RealQA [12]). To ensure a fair and rigorous comparison, we re-train these models on OA-Bench training set using their official implementations before evaluating their performance. For TAD66K, we also add comparisons against specialized models like TANet [6].

Evaluation Metrics. We evaluate performance using Spearman's Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC). Both metrics are calculated for the seen (versatility) and unseen (generalization) sets, and further broken down by deviation levels ("High," "Medium," "Low").

Implementation Details. We use Qwen2-VL-7B [34] as our backbone. Stage 1 utilizes LlamaFactory [47] with LoRA [7] (rank=128), trained for 2 epochs at learning rate (LR)=1e-4 with a batch size (BS) of 64. Stage 2 uses Verl [31] with LR=5e-7 (BS=16) and a rollout temperature of 0.6. We use the AdamW [19] optimizer.

6.2. Quantitative Analysis

Baseline Analysis. Vanilla MLLMs (e.g., GPT-4o, 0.517 overall PLCC) and reward models (e.g., HPSv3, 0.264 PLCC) show limited performance(Table 2). SOTA IAA baselines (e.g., RealQA, Q-align) reveal two key limitations: (1) *Versatility* (easy vs. hard): Performance on High themes is weaker than on Low (e.g., RealQA Unseen PLCC: 0.828→0.647), showing difficulty with a single criterion. (2) *Generalization* (seen vs. unseen): Performance drops on the Unseen Set (e.g., Q-align High PLCC: 0.690→0.605) indicating a generalization failure. These gaps confirm the OW-IAA challenge and validate OA-Bench.

Superiority of IAF. As shown in Table 2, IAF-Stage 1 surpasses RealQA, showing superior versatility (Seen High PLCC: 0.812 vs 0.710) and generalization (Unseen High PLCC: 0.716 vs 0.648). IAF-Stage 2 further boosts Unseen High PLCC to 0.749, establishing new SOTA.

Adaptation Boosts Versatility. Table 3 reveals a powerful "synergistic lift": After adapting on the "Unseen Set" (+ TAPA on unseen), the model's versatility on the "Seen Test Set" significantly increased (Low PLCC: 0.843 → 0.873). Notably, this gain surpassed the effect of adapting directly on the "Seen Set" (+ TAPA on seen) (Low PLCC 0.873 vs

Table 2. **Quantitative comparison on OA-Bench**, evaluating model Versatility (Seen) and Generalization (Unseen). Results are stratified by deviation levels (Low, Medium, High). For a fair comparison, all IAA Models are fine-tuned on the OA-Bench training set. Others (Vanilla MLLMs, Reward Models) are evaluated zero-shot. Note: IAF-Stage 2 used 1-shot unsupervised adaptation on Unseen themes only.

Method	Seen						Unseen						Overall	
	Low		Medium		High		Low		Medium		High			
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Vanilla MLLM														
GPT-4o[26]	0.558	0.554	0.524	0.499	0.554	0.531	0.516	0.470	0.542	0.508	0.556	0.540	0.542	0.517
Gemini-2.5 Pro[4]	0.470	0.429	0.498	0.438	0.434	0.412	0.467	0.378	0.483	0.396	0.469	0.395	0.470	0.408
Qwen2-VL-7B[34]	0.259	0.255	0.225	0.255	0.226	0.272	0.187	0.207	0.256	0.269	0.215	0.249	0.228	0.251
Qwen3-VL-8B[41]	0.324	0.351	0.347	0.366	0.347	0.357	0.304	0.326	0.372	0.387	0.376	0.403	0.345	0.365
Text-conditioned Reward Models														
HPSv3[22]	0.254	0.307	0.248	0.293	0.325	0.369	0.229	0.283	0.258	0.302	0.268	0.292	0.264	0.308
ImageReward[40]	0.236	0.234	0.236	0.248	0.348	0.377	0.233	0.247	0.259	0.263	0.282	0.296	0.266	0.277
UnifiedReward[35]	0.272	0.277	0.258	0.264	0.312	0.308	0.245	0.251	0.282	0.295	0.292	0.288	0.277	0.281
IAA Models														
Q-align[38]	0.806	0.806	0.775	0.769	0.673	0.690	0.797	0.774	0.742	0.712	0.621	0.605	0.736	0.726
Q-insight[13]	0.571	0.587	0.540	0.547	0.472	0.472	0.600	0.580	0.531	0.524	0.418	0.423	0.522	0.522
VisualQuality-R1[39]	0.654	0.653	0.576	0.572	0.538	0.493	0.614	0.599	0.563	0.558	0.449	0.438	0.566	0.552
RealQA[12]	0.818	0.837	0.786	0.805	0.681	0.710	0.816	0.828	0.757	0.770	0.647	0.648	0.751	0.766
IAF-Stage 1(*ours*)	0.829	0.843	0.813	0.831	0.788	0.812	0.800	0.817	0.772	0.785	0.708	0.716	0.785	0.801
IAF-Stage 2(*ours*)	**0.860**	**0.873**	**0.829**	**0.848**	**0.796**	**0.817**	**0.842**	**0.853**	**0.824**	**0.833**	**0.745**	**0.749**	**0.815**	**0.827**

Table 3. TAPA on Unseen Set Boosts Seen Set Performance.

Method	Low	Medium	High
IAF-Stage 1	0.843	0.831	0.812
+ TAPA on seen	0.866	0.847	0.812
+ TAPA on unseen	**0.873**	**0.848**	**0.817**

0.866).

We hypothesize this synergistic lift is an escape from overfitting, enabled by IAF's design. Although the ACI stage aims to induce a generalizable reasoning policy , the foundational policy, trained only on the seen set's incomplete view, inevitably learns "shortcuts" or gets trapped in a local optimum. When the model confronts unseen themes during TAPA, these shortcuts fail. This compels the model to activate and rely on the deeper, general aesthetic reasoning priors learned during ACI. Therefore, the "TAPA on unseen" process essentially **activates and refines the policy**, forcing it to escape the overfitting from the seen themes and converges to a better optimum.

Results on TAD66K. As shown in Table 4, we also achieved SOTA performance on the TAD66K benchmark. Our method (SRCC 0.526, PLCC 0.551) surpassed previous SOTA methods , including Q-align (SRCC 0.501, PLCC 0.531) and TANet (SRCC 0.513, PLCC 0.531), which was specifically designed for this task. Demenstrating the effectiveness of our IAF.

6.2.1. Ablation Studies

IAF Components. Table 5 validates component synergy. Solely adding CoT degrades performance (Unseen High PLCC 0.698→0.642), likely due to overemphasis on reasoning at the expense of scoring. Mixed training resolves this via generalizable reasoning priors, boosting High PLCC to 0.716. TAPA further adapts this to unseen criteria, achieving optimal performance across all levels (Low/Medium/High PLCC: 0.853/0.833/0.749) on the Unseen Test set.

Efficiency of TAPA. TAPA demonstrates efficiency across three dimensions: (1) *Data efficiency*: effective adaptation is achieved with few shots (1-5 samples per theme), with peak performance observed at 5 or 10 shots, depending on difficulty (Table 7); (2) *Training efficiency*: policy converges to near-optimal performance within 10 iterations (Table 6); (3) *Rollout efficiency*: model performance remains robust to sample size N, with computational overhead further reduced through sample reuse for GRPO updates (Table 8).

6.3. Qualitative Analysis

To demonstrate the **Versatility** and **Generalization** of IAF, we selected three representative images from the OA-Bench corresponding to the different themes, as shown in Fig. 4.

Table 4. Performance on TAD66K [6] benchmark.

Method	SRCC	PLCC
Q-align [38]	0.501	0.531
TANet [6]	0.513	0.531
PEAS [44]	0.415	0.444
Q-Instruct [37]	0.137	0.160
IAF (ours)	**0.526**	**0.551**

Table 5. **Ablation study of IAF components on Seen and Unseen Test Sets**. The components include: Direct (direct score prediction), CoT-Guided (CoT-guided prediction), and TAPA.

IAF Components			Seen			Unseen		
Direct	CoT-Guided	TAPA	Low	Medium	High	Low	Medium	High
✓	×	×	0.796	0.766	0.731	0.797	0.775	0.698
×	✓	×	0.790	0.768	0.741	0.749	0.731	0.642
✓	✓	×	0.843	0.831	0.812	0.817	0.785	0.716
✓	✓	✓	**0.866**	**0.847**	**0.812**	**0.853**	**0.833**	**0.749**

Table 6. Impact of train iterations on TAPA.

#Iter	Low	Medium	High
10	0.828	0.809	0.737
40	0.839	0.821	0.737
80	**0.853**	**0.833**	**0.749**
160	0.848	0.824	0.694

Table 7. Data efficiency of TAPA.

#Shots	Low	Medium	High
0	0.817	0.785	0.716
1	0.853	0.833	0.749
5	**0.855**	**0.835**	0.751
10	0.847	0.832	**0.754**

Table 8. Sensitivity to the rollout samples N.

N	Low	Medium	High
32	0.820	0.798	0.725
64	**0.828**	**0.809**	**0.737**
128	0.825	0.804	0.724

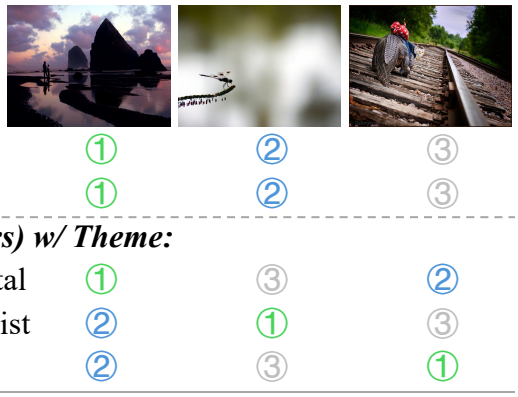

Figure 4. **Qualitative demonstration of IAF's Versatility and Generalization.** Number indicates the rank of the image. The CIAA model RealQA and Q-align both use a single, generic ranking. Our IAF model demonstrates Versatility by correctly ranking both a classic Theme (e.g., "Horizontal", i.e. "Horizontal Composition") and a non-classic Theme (e.g., "Minimalist"). Crucially, it demonstrates Generalization by successfully adapting to a user-customized Unseen Theme ("Unique," i.e., "Highly Unique Photography"), identifying the right image as Rank 1—the image the baseline ranked last.

CIAA models employ a single criterion, yielding a fixed ranking (left: Rank 1, right: Rank 3). In contrast, our IAF model demonstrates both **Versatility** and **Generalization**. It shows **Versatility** by correctly ranking both a classic theme ("Horizontal" e.g. "Horizontal Composition") and a non-classic theme ("Minimalist"), identifying the dragonfly (center) as Rank 1 for the latter.

Crucially, it demonstrates **Generalization** by successfully adapting to a user-customized unseen theme ("Unique," i.e., "Highly Unique Photography"). It captures the highly unique phenomenon of an alligator (the right image) with a backpack on the tracks, and identifies this image as the best, while SOTA CIAA models (RealQA and Q-align) deemed it the worst.

This analysis visually confirms IAF's ability to handle classic, non-classic, and user-customized aesthetic criteria, showing capabilities beyond the scope of CIAA models.

7. Conclusion

We introduced Open-World Image Aesthetic Assessment to address the pluralistic and open nature of real-world aesthetics, moving beyond the assumption of a single, static criterion. To tackle this challenge, we proposed the Induce-and-Adapt framework, which addresses these challenges through Aesthetic Criterion Induction via co-optimizing for direct score prediction and CoT-guided prediction, and Test-time Aesthetic Policy Adaptation through unsupervised refinement guided by crowd-simulated pseudo-labels. We further established OA-Bench, the first benchmark evaluating both versatility on seen themes and zero-shot generalization to unseen ones. Experiments demonstrate state-of-the-art performance on both OA-Bench and TAD66K, achieving a synergistic lift by improving generalization to unseen themes without sacrificing versatility on seen ones.

8. Limitations

While TAPA achieves high sample efficiency, this flexibility introduces a minor computational overhead during adaptation compared to static models. Additionally, IAF relies on textual descriptions to define themes; extending this to support richer, multi-modal theme definitions (e.g., few-shot visual examples) to enhance its expressive power presents a promising direction for future research.

References

[1] Challenging Technologies, LLC. DPChallenge - a digital photography contest. https://www.dpchallenge.com/, 2001–2025. Accessed: 2025-11-04. 5

[2] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 3

[3] Fei Gao, Yuhao Lin, Jiaqi Shi, Maoying Qiao, and Nannan Wang. Aesmamba: Universal image aesthetic assessment with state space models. In *ACM Multimedia 2024*, 2024. 2

[4] Google. Gemini: A Family of Highly Capable Multimodal Models. https://gemini.google.com/, 2024. Accessed: November 5, 2025. 5, 7

[5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 2, 5

[6] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 942–948. ijcai.org, 2022. 2, 3, 6, 8

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 10, 2021. 6

[8] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *ACM Multimedia 2024*, 2024. 2

[9] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pages 662–679. Springer, 2016. 2

[10] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910, 2020. 3

[11] Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi. Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[12] Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025. 1, 2, 6, 7

[13] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 7

[14] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. Transductive aesthetic preference propagation for personalized image aesthetics assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 896–904, 2022. 3

[15] Mingxiang Liao, Fang Wan, Yuan Yao, Zhenjun Han, Jialing Zou, Yuze Wang, Bailan Feng, Peng Yuan, and Qixiang Ye. End-to-end weakly supervised object detection with sparse proposal evolution. In *European conference on computer vision*, pages 210–226. Springer, 2022. 2

[16] Mingxiang Liao, Zonghao Guo, Yuze Wang, Peng Yuan, Bailan Feng, and Fang Wan. Attentionshift: Iteratively estimated part-based attention map for pointly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19519–19528, 2023.

[17] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

[18] Mingxiang Liao, Fang Wan, Zonghao Guo, and Qixiang Ye. Hierarchical attentionshift for pointly supervised instance segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[20] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*, pages 2206–2213, 2011. 3

[21] Pei Lv, Meng Wang, Yongbo Xu, Ze Peng, Junyi Sun, Shimei Su, Bing Zhou, and Mingliang Xu. Usar: An interactive user-specific aesthetic ranking framework for images. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1328–1336, 2018. 3

[22] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025. 6, 7

[23] Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. Lapis: A novel dataset for personalized image aesthetic assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6302–6311, 2025. 3

[24] Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, page 223–228. IEEE, 2020. 5

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 1, 2

[26] OpenAI. Chatgpt. https://chatgpt.com, 2025. Accessed: 2025-11-10. 6, 7

[27] Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The

photographer's eye: Teaching multimodal large language models to see, and critique like photographers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24807–24816, 2025. 2

[28] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision*, pages 638–647, 2017. 3

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5

[30] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8471–8480, 2021. 2

[31] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 6

[32] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 2

[33] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 5

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 6, 7

[35] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 7

[36] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision, 2023. 2

[37] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-instruct: Improving low-level visual abilities for multi-modality foundation models, 2023. 2, 8

[38] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi. 1, 2, 6, 7, 8

[39] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025. 7

[40] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 6, 7

[41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7

[42] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869, 2022. 3

[43] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L. Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22388–22397, 2023. 3

[44] Jooyeol Yun and Jaegul Choo. Scaling up personalized image aesthetic assessment via task vector arithmetic. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. 3, 8

[45] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021. 2

[46] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. 5

[47] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 6

[48] Haobin Zhong, Shuai He, Anlong Ming, and Huadong Ma. Rethinking personalized aesthetics assessment: Employing physique aesthetics assessment as an exemplification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2935–2944, 2025. 3

[49] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*, 2024. 2

[50] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long,

Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. 2