Unlocking the Essence of Beauty: Advanced Aesthetic Reasoning with Relative-Absolute Policy Optimization

Boyang Liu[1,4*] **Yifan Hu**[3,*] **Senjie Jin**[1,4,*] **Shihan Dou**[1] **Gonglei Shi**[4]
Jie Shao[4,†] **Tao Gui**[1,2,‡] **Xuanjing Huang**[1,2,‡]
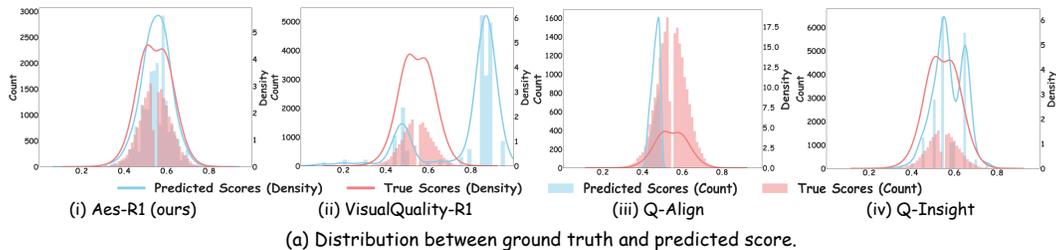
[1]Fudan University [2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Tsinghua University [4]Bytedance
{boyangliu25,sjjin24}@m.fudan.edu.cn tgui@fudan.edu.cn

Abstract

Advancements in multimodal large language models (MLLMs) demonstrate strong potential across various domains, with expanding applications in human-centered Image Aesthetic Assessment (IAA) through high-level cross modal understanding capacity. Compared to data-intensive supervised fine-tuning (SFT), the efficient and scalable reinforcement learning (RL) is a promising alternative for IAA. However, directly applying RL to IAA fails to inspire reasoning patterns without prior aesthetic knowledge, and modeling an appropriate reward proxy to evaluate aesthetic scores is intricate. To this end, we propose Aes-R1, a comprehensive IAA framework. Concretely, Aes-R1 integrates a data pipeline, AesCoT, to construct and filter high-quality aesthetic reasoning data used for cold-start. After teaching the model to generate structured explanations before scoring, we employ the Relative-Absolute Policy Optimization (RAPO), a novel RL algorithm that jointly optimizes absolute score regression and relative ranking order, improving both per-image accuracy and cross-image preference judgments. Extensive experiments demonstrate that Aes-R1 improves the backbone's average PLCC/SRCC by 47.9%/34.8%, surpassing state-of-the-art baselines of similar training size. More ablation studies validate Aes-R1's robust generalization in out-of-distribution scenarios with limited supervision.[1]

(a) Distribution between ground truth and predicted score.

(b) Comparison of various methods.

Figure 1: (a) Distributions of ground truth and predictions on the AVA test set. Aes-R1 achieves the closest alignment with the ground truth compared to baseline methods. (b) The case study shows that Aes-R1 surpasses other methods, which suffer from prediction errors and limited ordinal discernment, based on the relative-absolute reward design.

---

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) (Qwen et al., 2025; OpenAI et al., 2024; Team et al., 2025) have shown remarkable potential in image assessment (Xu et al., 2023; Wu et al., 2023b). To date, most efforts center on image quality assessment (IQA) (Wu et al., 2024; 2023a; You et al., 2025), which measures images' sharpness and clarity. As the development of creative design and social media recommendations, there is an emerging demand for reliable metrics that go far beyond pixel-level fidelity and some researchers have paid attention to image aesthetic assessment (IAA) (Huang et al., 2024; Zhou et al., 2024), leveraging the cross-modal understanding capabilities of MLLMs to capture high-level features such as texture, color harmony and emotional impact, broadening vision tasks to the interpretable and human-centered aesthetic modeling.

Current IAA methods generally adopt supervised fine-tuning (SFT) on score-based targets (Wu et al., 2023a; Cao et al., 2025). While such implicit reasoning restricts MLLMs from aligning visual elements with multidimensional aesthetic criteria and lacks explainability in practice (Wu et al., 2025; Li et al., 2025), some propose generating reasoning prior to the answer, supporting the integration of fine-grained descriptive feedback with aesthetic scores (Li et al., 2025). Nonetheless, existing resources (He et al., 2022; Yang et al., 2022) primarily consist of score-based data, and artist-level rationales necessitate expensive expert annotation. Moreover, as presented in Tab. 3, extensive SFT causes the model to learn bias and overfit on curated dataset, rapidly reducing entropy and hindering further optimization, making it costly and difficult to scale.

Reinforcement learning (RL) is an effective alternative and has shown strong efficiency and generalization in goal-driven tasks (Chu et al., 2025a; Chen et al., 2025a). However, in our pilot study, directly applying RL to IAA exposes several challenges. 1) Although achieving high performance (Tab. 3, SFT 0 epoch, RAPO), the end-to-end RL fails to activate aesthetic reasoning patterns in the backbone due to the absence of corresponding corpus pre-training (Gandhi et al., 2025; Ji et al., 2025), resulting in the poor explanatory rationales as shown in Fig. 6 (AesR1-Zero). 2) Modeling an appropriate reward proxy to evaluate aesthetic score is intricate since the absence of unified criteria for aesthetic preferences. We reproduce the related works on AVA (Murray et al., 2012) and find that, although VisualQuality-R1 (ii) (Li et al., 2025) successfully separates high- and low-quality images using the rank reward, it fails to calibrate scores, while Q-insight (iv) (Wu et al., 2025) suffers from peak mismatch when optimizing with a scalar reward, as presented in Fig. 1.

Integrating aesthetic priors and an effective reward mechanism is crucial for IAA in RL optimization. Inspired by Deepseek-R1 (Guo et al., 2025a), we find that cold-start bootstrapping with thousands of high quality reasoning data effectively induces robust aesthetic analysis patterns (Gandhi et al., 2025; Shao et al., 2025). Following this, RL transforms IAA into a data efficient self-learning process, primarily guided by the reward signal. Building on the results in Fig. 1, we observe that human aesthetic judgment is inherently context-dependent besides absolute quality evaluation, as also noted by Immanuel Kant's assertion that "*there can be no objective rule of taste which shall determine by means of concepts what is beautiful*" (Kant, 1790). Therefore, we argue the IAA reward should comprise two complementary facets: a relative dimension that ranks images by comparative appeal and an absolute dimension that assesses intrinsic aesthetic merit, thereby simultaneously enforces ordinal consistency and calibrates absolute aesthetic scores.

Building on the above insights, we propose **Aes-R1** framework. We first construct an automatic data pipeline **AesCoT**, which effectively generates high-quality aesthetic reasoning data along five dimensions based on original image-score pairs. Warming up on this data equips MLLMs with aesthetic cognitive behaviors that generating reliable explanations before the aesthetic scores. We then utilize Relative-Absolute Policy Optimization (RAPO), a novel RL algorithm that optimizes both absolute and relative aesthetic preferences. Specifically, the model's gradient updates are simultaneously steered by the absolute-error reward, calibrating aesthetic score constraints and relative-rank reward, aligning pairwise ranking consistency, which greatly improves effectiveness and stability in the exploration and exploitation process. Empirically, trained with only 15K data, Aes-R1 surpasses the state-of-the-art baselines of comparable size with the improvements on backbones by 47.9%/34.8% PLCC/SRCC. The distribution alignment experiment (see Fig. 1, a) and reward ablations (see Sec. 4.3) demonstrate the effectiveness of our RAPO as well as generalization in the OOD setting (see Sec. 4.2). We further analyze the training recipe of SFT and RL from the entropy perspective (Cui et al., 2025b) (see Sec. 4.3).

In summary, our main contributions can be summarized as follows:

❶ We construct the effective and automatic data pipeline, **AesCoT**, which builds high-quality aesthetic reasoning data along five dimensions based on original image-score pairs, significantly reducing labor cost. We will release our **AesCoT-3K** and **AesCoT-10K** to facilitate future research in IAA.

❷ We develop a two-stage training pipeline, **Aes-R1**, based on Relative-Absolute Policy Optimization (RAPO) that jointly optimizes absolute score regression and relative ranking, enabling balanced modeling of intrinsic merit and comparative preference in IAA.

❸ We conducted extensive experiments to demonstrate the performance, robustness and generalization of Aes-R1, which achieves state-of-the-art performance across multiple IAA benchmarks with only 15K training samples. We also analyze the reward design and training recipe for SFT and RL, offering valuable insights.

2 RELATED WORK

Score-Based Image Aesthetic Assessment. Research on Image Aesthetic Assessment (IAA) has progressed from rule-based and handcrafted pipelines (Bhattacharya et al., 2010; Mavridaki & Mezaris, 2015; Mittal et al., 2013; 2012) to deep models supervised by human-annotated mean opinion scores (MOSs). Deep learning IAA models (Talebi & Milanfar, 2018; Ke et al., 2021; Ghosal & Smolic, 2022; He et al., 2022; Behrad et al., 2025) became dominant after large-scale aesthetic datasets (Murray et al., 2012; He et al., 2022; Ren et al., 2017; Yang et al., 2022; Kong et al., 2016) enabled direct pixel-level representation learning via powerful networks. More recently, pretrained MLLMs have demonstrated remarkable cross-modal aesthetic perception and evaluation through visual-language pretraining (Ke et al., 2023), prompt engineering (Zhu et al., 2024; Sheng et al., 2025), and supervised fine-tuning (Zhang et al., 2023; Cao et al., 2025; Liao et al., 2025; Zhou et al., 2024). Nevertheless, score-only supervision without explicit linguistic rationales limits aesthetic understanding, weakens generalization and requires redundant dataset-specific training.

Reinforcement Learning-Enhanced Model-as-a-Judge. Model-as-a-Judge approach serves as a novel and broadly applicable paradigm for addressing open-ended evaluation problems in the era of Large Models (Gu et al., 2025). Prior works on Model-as-a-Judge (Zheng et al., 2023; Zhu et al., 2025; Chen et al., 2024; Lee et al., 2024) as well as the Generative Reward Models (GRMs) (Mahan et al., 2024) highlight that reasoning step-by-step before verdict improves the stability and consistency (Kim et al., 2024; Saha et al., 2025). Capitalizing on recent advances in post-training (Liu et al., 2025; Yu et al., 2025), various large reasoning judge models have emerged (Liu et al., 2025; Chen et al., 2025b; Whitehouse et al., 2025) and have become popular fro automated assessment. Extending to image evaluation, researchers move beyond scalar scoring (Wu et al., 2023a; You et al., 2025; Zhou et al., 2024) by applying reinforcement learning to image quality assessment via step-by-step reasoning (Li et al., 2025; Wu et al., 2025). However, these methods are confined to either binary outcome supervision or rank supervision and have yet to be extended to aesthetic assessment, which demands finer-grained perceptual and semantic understanding capability.

3 METHODOLOGY

There is indeed an absence of a principled visual reasoning framework for IAA, which contributes to the scarcity of high-quality aesthetic reasoning data and the misalignment of single-objective optimization with nuanced human preferences. To address this, we propose our Aes-R1 framework. Specifically, we introduce (i) AesCoT, an aesthetic reasoning data construction pipeline that automatically curates diverse, reliable annotations with multi-dimensional aesthetic rationales. (ii) A two-stage training pipeline based on Relative-Absolute Policy Optimization (RAPO) that jointly optimizes pairwise preferences and calibrated absolute scores under a unified objective.

3.1 PROBLEM FORMULATION

Given an image dataset that contains M image-score pairs $\mathcal{D} = \{(\mathcal{I}_i, s_i), i \in [1, M]\}$, considering a pretrained MLLM with parameter θ and policy $\pi_\theta(\cdot | \mathcal{I}, \mathcal{P})$, we aim to fine-tune the model for ideal
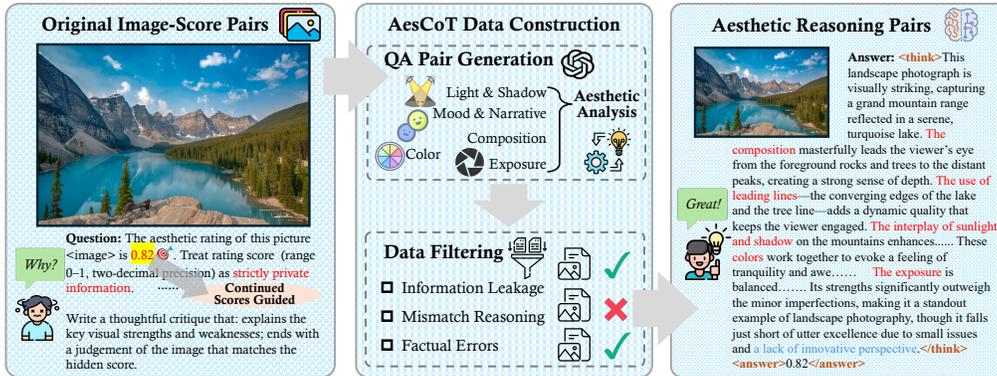
Figure 2: Overview of AesCoT construction pipeline. Starting from original image-score pairs, we mask the continuous aesthetic score and prompt experts to produce CoT critiques along five aesthetic dimensions. Automated checks and human audits then remove any score leakage, reasoning–score mismatch, or factual errors, yielding high-quality, interpretable multimodal aesthetic reasoning data.

aesthetic reasoning that provides the comprehensive explanation c and an overall aesthetic score s for every image in dataset $\mathcal{I}$ under prompt $\mathcal{P}$:

$$\tau = (c, s) \sim \pi_\theta(\cdot | \mathcal{I}, \mathcal{P}) \tag{1}$$

We utilize policy gradient methods (Sutton et al., 1999) to optimize the MLLM by maximizing the expected cumulative reward $R(\tau)$ on sampled aesthetic reasoning trajectories τ: $\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot | \mathcal{I}, \mathcal{P})}[R(\tau)]$. By performing gradient ascent on $\mathcal{J}(\theta)$:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot | \mathcal{I}, \mathcal{P})} \left[R(\tau) \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(a_t \mid s_t\right) \right], \tag{2}$$

where a_t is selected from the vocabulary space $\mathcal{V}$ and $s_t = (\mathcal{I}, \mathcal{P}, a_{<t})$ is the state at time step t.

3.2 AESCOT

Most existing datasets (e.g., AVA (Murray et al., 2012)) annotate image aesthetics using direct Mean Opinion Scores (MOSs). Chain-of-Thought (CoT) (Wei et al., 2022) has been shown to substantially enhance the performance and interpretability by breaking down the reasoning process into intermediate steps (Bi et al., 2025). However, high-quality aesthetic reasoning trajectories require artist-level curation and iterative refinement, making such data rare and costly to scale. Distillation from a strong model is an efficient and scalable approach (Guo et al., 2025a) and was validated in domains from math reasoning (Jie et al., 2023) to visual question answering (Shao et al., 2024a). Based on these, we propose AesCoT, to our knowledge, the first automatic pipeline for aesthetics reasoning data construction.

We display the construction pipeline of AesCoT (see Fig. 2). For every collected original image-score pair $(\mathcal{I}_i, s_i)$ in $\mathcal{D}$, we first intentionally conceal the continuous score from outputs and enable the powerful close-sourced MLLM to give aesthetic analysis over five aesthetic dimensions (see Appendix E.1) (light and shadow, mood and narrative, composition, color, exposure) that match the aesthetic levels, which are divided into bad (0-0.4), fair (0.4-0.7) and good (0.7-1.0) to emphasize aesthetic differences. Then we concatenate the aesthetic analysis with the ground truth score, obtaining $\mathcal{D}_{CoT} = \{(\mathcal{P}, \mathcal{I}_i, c_i, s_i), i \in [1, M]\}$ where c_i is the reasoning trajectory towards s_i. Finally, we carefully filter the outputs to eliminate score leakage, reasoning inconsistencies, and other factual errors through automated checks and human audits, generating a reliable aesthetic reasoning dataset $\mathcal{D}_{AesCoT} = \mathcal{F}(\mathcal{D}_{CoT}) = \{(\mathcal{P}, \mathcal{I}_i, c_i, s_i) \mid \|\mathcal{E}_i\| = 0\}$, where $\mathcal{F}$ is the data filter process and $\mathcal{E}_i = (e_i^{\text{leak}}, e_i^{\text{align}}, e_i^{\text{fact}})$ represent different error types. Dataset details can be found in **??**.

3.3 RELATIVE-ABSOLUTE POLICY OPTIMIZATION (RAPO)

Aesthetic evaluation is inherently subjective, shaped by individual preferences, cultural experience, and contextual cues. In real-world settings, human judgments are typically expressed through com-
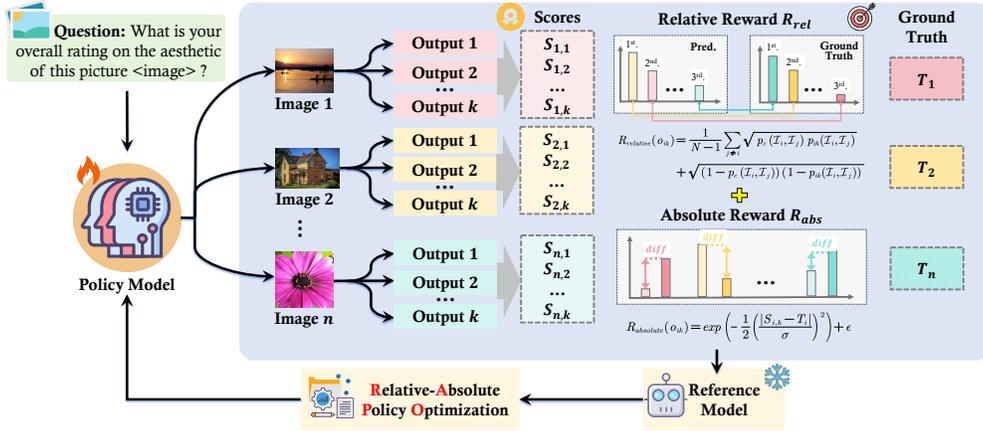
Figure 3: The training pipeline of Relative Absolute Policy Optimization (RAPO). Given a batch of training images, the policy model generates multiple outputs for aesthetic rating questions. RAPO computes both pairwise ranking based relative reward and score regression error based absolute reward for each output, then optimizes the policy model to align with human aesthetic preferences.

parative assessments, in addition to direct absolute scores, with judgments manifesting along two complementary axes: relative preference (which image is favored) and absolute magnitude (by how much). In view of this, we introduce Relative Absolute Policy Optimization (**RAPO**), a novel reinforcement learning algorithm that integrates relative assessment learning with absolute numerical optimization to better align model predictions with human aesthetic preferences.

Fig. 3 describes the training pipeline of RAPO. For a mini training batch $\{(\mathcal{I}_1, s_1), \ldots, (\mathcal{I}_N, s_N)\}$, where the training size is N, we follow GRPO (Shao et al., 2024b) to sample a group of K outputs $\{o_{i1}, o_{i2}, \ldots, o_{iK}\}$ for every prompt-image input $(\mathcal{P}, \mathcal{I}_i)$ from the old policy π_θ^{old}.

A robust and reliable reward signal is significant for aligning MLLMs with human aesthetic preferences. To this end, RAPO jointly optimizes the relative rank preference as well as the absolute score regression error.

Relative Rank Reward.   Learning to rank is a crucial paradigm for modeling human preference (Burges et al., 2005; Xia et al., 2008), and has proven effective for perceptual quality and aesthetic assessment (Liu et al., 2017; Talebi & Milanfar, 2018). Following (Tsai et al., 2007; Sun et al., 2024; Wu et al., 2025; Zhang et al., 2023; You et al., 2025), our relative rank reward based on FRank is bounded, continuous, differentiable, and aligns directly with pairwise ranking consistency, and thus provides stable, precise signals for optimizing pairwise image aesthetic preference.

Specifically, we assume that the aesthetic score of an image follows the Gaussian distribution $s \sim \mathcal{N}(\mu, \sigma^2)$. Consequently, the score difference $s_i - s_j$ follows $\mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$. Let o_{ik} denote the k-th predicted score for image i, the pairwise comparison probability against image j should be

$$p_{ik}(\mathcal{I}_i, \mathcal{I}_j) = \Phi\left(\frac{o_{ik} - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 + \gamma}}\right), i \neq j, \tag{3}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We estimate μ_j as the mean of k predicted scores of image j, and set a small positive constant γ to avoid division-by-zero.

$$r_{rank}(o_{ik}) = \frac{1}{N-1} \sum_{j \neq i} \sqrt{p_c(\mathcal{I}_i, \mathcal{I}_j) p_{ik}(\mathcal{I}_i, \mathcal{I}_j)} + \sqrt{(1 - p_c(\mathcal{I}_i, \mathcal{I}_j))(1 - p_{ik}(\mathcal{I}_i, \mathcal{I}_j))}, \tag{4}$$

where p_c is a binary preference label based on ground-truth MOSs of two images

$$p_c(\mathcal{I}_i, \mathcal{I}_j) = \begin{cases} 1, & s_i \geq s_j \\ 0, & s_i < s_j \end{cases}. \tag{5}$$

Absolute Error Reward. Although the rank reward reinforces the model to order images correctly, it leaves the calibrated scoring underconstrained. To fix this, we incorporate a continuous absolute-error reward that explicitly calibrates predicted scores toward the ground-truth MOSs, where σ is a hyperparameter and ϵ is a small positive constant to avoid reward being zero.

$$r_{abs}(o_{ik}) = exp\left(-\frac{1}{2}\left(\frac{|o_{ik} - s_i|}{\sigma}\right)^2\right) + \epsilon. \tag{6}$$

The total reward function of RAPO is combined with relative rank reward and absolute error reward

$$r = r_{rank} + r_{abs}. \tag{7}$$

Training Aes-R1 for Aesthetic Reasoning. To fully inspire the aesthetic reasoning capability of the MLLM, we first leverage the constructed AesCoT dataset to fine tune the model as a cold start stage, optimizing the negative log-likelihood of the aesthetic reasoning trajectory and the final score

$$\mathcal{L}_{sft}(\theta) = \mathbb{E}_{(\mathcal{P},\mathcal{I},c,s)\sim\mathcal{D}_{CoT}}\left[-\log\pi_\theta(c,s|\mathcal{P},\mathcal{I})\right]. \tag{8}$$

For the prompt-image pair $(\mathcal{P},\mathcal{I}_i)$, we calculate the rewards of all sampled outputs, yielding rewards $R_i = \{r_1, r_2, \ldots, r_K\}$. The advantage $\hat{A}_{k,t}$ for each token time step t in the k-th output is given by subtracting the group average and dividing by the standard deviation

$$\hat{A}_{k,t} = \frac{r_k - \mu(R_i)}{\sigma(R_i)}. \tag{9}$$

Following DAPO (Yu et al., 2025), we adopt a higher clipping threshold and reduce the KL penalty terms. RAPO optimizes the policy by maximizing the following objective

$$\mathcal{J}_{\text{RAPO}}(\theta) = \mathbb{E}_{(\mathcal{P},\mathcal{I})\sim Q,\, o_i\sim\pi_\theta^{\text{old}}(\cdot|\mathcal{P},\mathcal{I})}$$

$$\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|o_{ik}|}\sum_{t=1}^{|o_{ik}|}\min\left(r_{k,t}(\theta)\hat{A}_{k,t},\,\text{clip}\left(r_{k,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_{k,t}\right) - \beta\,\mathcal{D}_{\text{KL}}(\pi_\theta\,\|\,\pi_{\text{ref}})\right], \tag{10}$$

where π_{ref} denotes the reference policy of pretrained model, and the probability ratio is defined as

$$r_{k,t} = \frac{\pi_\theta(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})}{\pi_\theta^{old}(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})}, \tag{11}$$

and the KL penalty is approximated by

$$\mathcal{D}_{\text{KL}}(\pi_\theta\,\|\,\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})}{\pi_\theta(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})} - \log\frac{\pi_{\text{ref}}(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})}{\pi_\theta(o_{k,t}|\mathcal{P},\mathcal{I},o_{k,<t})} - 1. \tag{12}$$

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Dataset and Metrics. We adopt five open-source general image aesthetic assessment datasets: **TAD66K** (He et al., 2022), **AVA** (Murray et al., 2012), **FLICKR-AES** (Ren et al., 2017), **PARA** (Yang et al., 2022), and **AADB** (Kong et al., 2016). The model is trained on TAD66K, AVA, and Flickr-aes, with PARA and AADB serving as out-of-distribution (OOD) datasets. Mean Opinion Scores (MOSs) across these datasets are normalized into the range $[0, 1]$. For evaluation, model performance is quantified using the Pearson linear correlation coefficient (**PLCC**) and the Spearman rank-order correlation coefficient (**SRCC**) between predicted scores and ground-truth annotations. The implementation details can be found in Appendix C.

Baselines and Implementation. We compare our method with : (1) popular LLMs with various sizes, capacities and reasoning modes, including **Qwen-2.5-VL** series (Qwen et al., 2025) , **Seed-1.5-VL-thinking** (Guo et al., 2025b), **GPT4o** (OpenAI et al., 2024) and **GPT4.1** (OpenAI, 2025); (2) HandCrafted methods like **NIQE** (Mittal et al., 2013) and **BRISQUE** (Mittal et al., 2012); (3) Deep Learning based methods including **MUSIQ** (Ke et al., 2021) and **NIMA** (Talebi & Milanfar, 2018); (4) MLLM based methods including **VILA** (Ke et al., 2023), **Laion-Aes** (Schuhmann et al., 2022), **Q-Align** (Wu et al., 2023a), **ArtiMuse** (Cao et al., 2025), **DeQA-Score** (You et al., 2025), **Q-Insight** (Li et al., 2025) and **VisualQuality-R1** (Wu et al., 2025). The implementation details is in Appendix C.

Model	TAD66K		AVA		FLICKR		PARA		AADB		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Vanilla MLLM												
Qwen2.5-VL-7B	0.2282	0.2242	0.3518	0.3684	0.5179	0.5696	0.5966	0.6252	0.4480	0.5076	0.4285	0.4589
Qwen2.5-VL-32B	0.2300	0.2715	0.3893	0.4170	0.5800	0.6527	0.7176	0.7118	0.5060	0.5138	0.4846	0.5134
Seed1.5-VL-thinking	0.2908	0.2974	0.4533	0.4871	0.5837	0.6374	0.6134	0.6364	<u>0.5279</u>	**0.5577**	0.4938	0.5232
GPT4o-2024-08-06	0.2686	0.2979	0.4687	0.4939	0.5469	0.5507	0.6916	0.6719	0.5209	0.5314	0.4993	0.5092
GPT-4.1-2025-04-14	0.2707	0.3298	0.4846	0.5594	0.5909	0.6062	0.7290	0.7203	0.5102	0.5296	0.5171	0.5491
Hand-Crafted												
NIQE (2013)	0.1332	0.0811	0.1226	0.0448	0.1024	0.0487	0.3546	0.2478	0.0698	0.0566	0.1565	0.0958
BRISQUE (2012)	0.0294	0.0243	0.0310	0.0360	0.0830	0.0910	0.1288	0.0783	0.0565	0.0558	0.0657	0.0571
Deep-Learning Based												
NIMA (2018)	0.3885	0.3654	<u>0.6120</u>	<u>0.6361</u>	0.5130	0.4796	0.5868	0.5709	0.3886	0.3904	0.4978	0.4885
MLLM Based												
Q-Align* (2023a)	0.3498	0.3627	0.5215	0.5407	0.6231	0.6473	0.6181	0.6262	0.4474	0.4508	0.5120	0.5255
DeQA-Score* (2025)	<u>0.3985</u>	0.3885	0.5718	0.5896	<u>0.7057</u>	<u>0.6889</u>	0.7234	0.6796	0.4859	0.4948	0.5771	0.5663
Q-Insight* (2025)	0.3980	0.3886	0.5964	0.5898	0.7012	0.6769	<u>0.7745</u>	0.7428	0.5069	0.5184	<u>0.5954</u>	0.5813
VisualQuality-R1* (2025)	0.3082	<u>0.3915</u>	0.4407	0.6195	0.5524	0.6769	0.5376	<u>0.7507</u>	0.3754	0.5363	0.4429	<u>0.5930</u>
Aes-R1	**0.4513**	**0.4248**	**0.6702**	**0.6619**	**0.7243**	**0.6973**	**0.7842**	**0.7666**	**0.5386**	<u>0.5423</u>	**0.6337**	**0.6186**

Table 1: PLCC and SRCC results compared to vanilla MLLM, hand-crafted, deep-learning based and MLLM based methods. *Results are retrained in our 15K combined train set. (DeQA-Score is trained on AVA and Flickr-aes in combination, owing to the absence of per-image standard deviation data in TAD66K). The best results are highlighted in bold, and the runner-ups are underlined.
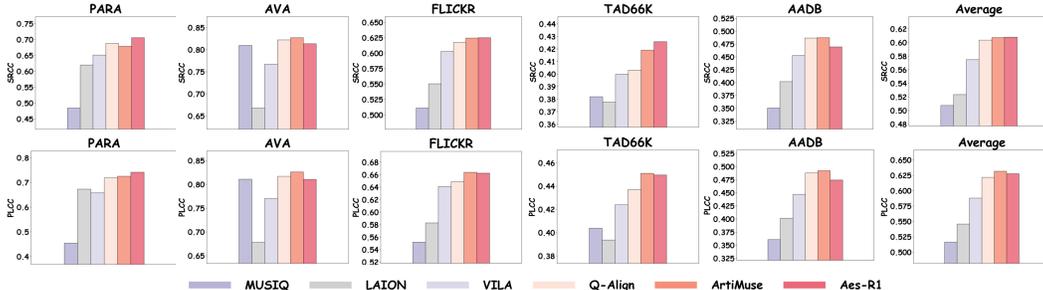

Figure 4: PLCC and SRCC results of models trained only on AVA dataset to compare the generalization ability. Aes-R1 is able to achieve performance comparable to models such as ArtiMuse (Cao et al., 2025), which is pretrained on large-scale human-annotated corpus.

4.2 QUANTITATIVE RESULT

Main Results. We adopt a combined training set of 15K samples randomly collected from the official train set of AVA, TAD66K and FLICKR-AES in a 2:2:1 ratio. Tab. 1 indicates several findings: 1) Aes-R1 significantly outperforms all baseline methods on in-domain datasets, achieving the highest average PLCC (0.6337) and SRCC (0.6186) scores across five benchmarks, with improvements of over 0.2 in both PLCC and SRCC compared to the backbone. 2) Compared to SFT-based methods, the efficient and scalable RL-based methods demonstrate superior performance, highlighting the promise of our Aes-R1 framework in IAA. 3) Although VisualQuality-R1, trained on the same data, successfully separates high- and low-quality images using rank reward (as shown in Fig. 1) and achieves a high SRCC (0.5825), it fails to calibrate scores to the empirical distribution (PLCC 0.4429). This highlights the rationale behind Aes-R1's relative-absolute reward design, which achieves the closest alignment with the ground-truth distribution, featuring a well-matched peak and reduced tail deviation. We further conduct reward ablations in Sec. 4.3.

Generalization Ability. In Tab. 1, Aes-R1 attains strong OOD performance on PARA and AADB trained on combined datasets compared to MLLM based baselines. We further validate Aes-R1's generalization in single dataset setting by training Qwen-2.5-VL Instruct on the AVA train set and compare it with the official baseline checkpoints on the test set. As shown in Fig. 4, Aes-R1 achieves the highest SRCC and the second-highest PLCC on average. Note that ArtiMuse is pretrained on a large-scale human-annotated corpus. The results indicate RL with appropriate feedback can achieve strong cross-task generalization without the need for manually annotated data. Detailed results can be found in Appendix D.1.

Binary reward	Error reward	Rank reward	TAD66K		AVA		FLICKR		PARA		AADB		Average	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
✓	×	×	0.2286	0.2414	0.3872	0.4120	0.5020	0.5336	0.6109	0.6162	0.3989	0.4133	0.4255	0.4433
×	✓	×	0.3981	0.3872	0.5928	0.5895	0.7019	0.6829	0.7034	0.7023	0.4313	0.4379	0.5655	0.5600
×	×	✓	0.3170	0.3638	0.5098	0.5956	0.5550	0.6945	0.5171	0.7582	0.3719	0.5421	0.4542	0.5908
✓	×	✓	0.3884	0.3785	0.5963	0.5932	0.6996	0.6780	0.7694	0.7452	0.5284	0.5176	0.5964	0.5825
×	✓	✓	0.4147	0.3915	0.6294	0.6242	0.7422	0.7173	0.8192	0.7813	0.5428	0.5369	0.6297	0.6102

Table 2: Ablation studies of different reward combinations. Our RAPO proposed error-rank reward significantly outperforms others.

SFT(epoch)	RL	Average Entropy	TAD66K		AVA		FLICKR		PARA		AADB		Average	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
0	N/A	0.921	0.2282	0.2242	0.3518	0.3684	0.5179	0.5696	0.5966	0.6252	0.4480	0.5076	0.4285	0.5490
0	RAPO	0.961	0.4147	0.3915	0.6294	0.6242	0.7422	0.7173	0.8192	0.7813	0.5428	0.5369	0.6297	0.6102
1	N/A	1.609	0.2570	0.2548	0.3755	0.3734	0.5032	0.4731	0.5561	0.5989	0.3566	0.3416	0.4097	0.4084
1	RAPO	1.626	0.4513	0.4248	0.6702	0.6619	0.7243	0.6973	0.7842	0.7666	0.5386	0.5423	0.6337	0.6186
2	N/A	1.417	0.3215	0.3214	0.4652	0.4602	0.5209	0.5170	0.5941	0.6377	0.4104	0.4164	0.4624	0.4705
2	RAPO	1.391	0.4460	0.4326	0.6592	0.6568	0.6822	0.6667	0.7777	0.7290	0.4486	0.4663	0.6027	0.5903
10	N/A	0.705	0.3323	0.3248	0.4752	0.4730	0.5745	0.5594	0.6045	0.6589	0.4369	0.4354	0.4847	0.4903
10	RAPO	0.716	0.3215	0.3214	0.4652	0.4602	0.5209	0.5170	0.5941	0.6377	0.4104	0.4164	0.4624	0.4705

Table 3: The performance when RAPO is initialized from SFT checkpoints (0, 1, 2, 10 epochs), together with the average token entropy of the starting policy. Moderate SFT maximizes downstream performance, while excessive SFT declines entropy and RL gains, hindering OOD performance.

4.3 ABLATION STUDY

Reward Function Analysis. We conduct ablation studies on the reward function design, the key module of our RAPO algorithm. We investigate three kinds of reward designs: **binary reward**, **relative rank reward** (Eq. 4) and **absolute error reward** (Eq. 6). The binary reward was used in (Li et al., 2025), which only supervises the outcome as

$$r_{binary} = \begin{cases} 1, & |s_{pred} - s_{gt}| < \epsilon \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

We train the backbone on the 15K combined dataset via direct reinforcement learning without the cold-start stage to analyze the reward configurations. As summarized in Tab. 2, single-objective rewards exhibit distinct behaviors. The error reward markedly outperforms the binary reward, as it provides a continuous and precise reward signal that aligns with the score distribution. Additionally, the rank-only reward optimizes relative ordering effectively, yielding high SRCC but depressed PLCC, which matches the results in Tab. 1, reflecting the imbalance in predicting aesthetic scores.

Compared with single-objective rewards, the combining objectives improves the generalization capability, resulting in stronger performance in the out-of-distribution (OOD) datasets. Notably, the Aes-R1 configuration (Error- Rank) achieves best average PLCC and SRCC across all benchmarks.

Analysis of SFT and RL. We first attempt RAPO training without SFT on the AesCoT dataset (Aes-R1-Zero) Guo et al. (2025a). As illustrated in Fig. 6, Aes-R1-Zero achieves reasonable scoring accuracy but generates superficial and generic explanations, posing potential reward-hacking risks (Gao et al., 2023). This indicates the model lacks sufficient aesthetic prior, while recent studies (Chu et al., 2025b) highlight the importance of SFT in providing a strong prior and stabilizing output formats. Consequently, Aes-R1 framework consists of two training stages, SFT and RL. To find the reasonable allocation between them, we conduct initialization experiments. As shown in Tab. 3, the average PLCC/SRCC increased significantly from 0.4097/0.4084 to 0.4847/0.4903 while the average token entropy decreased from 1.609 to 0.716 after 1 epoch, 2 epochs and 10 epochs SFT. We then train RAPO based on above SFT checkpoints.

On the RL side, maintaining appropriately high entropy can foster exploration, stabilize training and mitigate entropy collapse (Cui et al., 2025a; Wang et al., 2025a; Cheng et al., 2025), which is consistent with the RAPO results, where checkpoints with lower performance but higher entropy exhibit significant improvements. Excessive SFT leads to overfitting on the data and narrows the policy exploration, resulting in only marginal gains after RL. Therefore, in this paper, we set 1 epoch cold-start, which achieves the best scoring performance of 0.6337/0.6186 during reinforcement learning, as well as expert-level aesthetic analysis (Aes-R1 in Fig. 6).

Relative	Absolute	TAD66K		AVA		FLICKR		PARA		AADB		Average	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
1.0	0	0.3170	0.3638	0.5098	0.5956	0.555	0.6945	0.5171	0.7582	0.3719	0.5421	0.4542	0.5908
0.8	0.2	0.4059	**0.4263**	0.5596	**0.6464**	0.6318	**0.7596**	0.6752	**0.8392**	0.5041	**0.5966**	0.5553	**0.6536**
0.6	0.4	0.3986	0.4222	0.6274	0.6317	0.7075	0.7328	0.7750	0.7993	0.5305	0.5422	0.6078	0.6256
0.5	0.5	0.4147	0.3915	0.6294	0.6242	**0.7422**	0.7173	0.8192	0.7813	0.5428	0.5369	**0.6297**	0.6102
0.4	0.6	**0.4282**	0.4082	0.6478	0.6234	0.7254	0.6984	0.7914	0.7660	**0.5432**	0.5278	0.6272	0.6048
0.2	0.8	0.4251	0.3885	**0.6492**	0.6392	0.7280	0.6990	0.7764	0.7635	0.5291	0.5185	0.6216	0.6017
0	1.0	0.3981	0.3872	0.5928	0.5895	0.7019	0.6829	0.7034	0.7023	0.4313	0.4379	0.5655	0.5600

Table 4: The performance when changing weighting coefficient of the RAPO reward.

In contrast, extending SFT training to 2 or up to 10 epochs leads to depressed entropy and limited exploration, and initializing RAPO results in worse performance and poorer generalization capability. Finally, we choose to train Aes-R1 with only 1 epoch cold-start to cultivate aesthetic reasoning patterns, followed by RAPO training that inspires effective self-exploration. Compared with methods that rely on large volumes of constructed data, our Aes-R1 training pipeline achieves strong results while overcoming the prohibitive cost of data collection and curation. Detailed analysis of entropy can be found in Appendix D.2.

Analysis of Weighting Coefficients in RAPO To better understand the interplay between the relative rank reward and the absolute error reward, we conduct ablation experiments. We change the weighting coefficients in the RAPO algorithm to explore the mutual relationship between the two reward components. The reward function can be written as:

$$r = \lambda_{rank} r_{rank} + \lambda_{abs} r_{abs}. \tag{14}$$

As shown in Tab. 4, we retrain the model using different reward configurations. More intuitively, we visualize the trends of the PLCC and SRCC metrics as a function of the weighting coefficients in Fig. 5. The results reveal several compelling observations: When relying solely on the relative rank or absolute error reward ($\lambda_{rank} = 1.0$ or $\lambda_{abs} = 1.0$), the model achieves significant performance degradation. Besides, when both applying relative and absolute reward, it's worth noting and compelling that the two reward components exhibit complementary characteristics. As shown in Fig. 5, PLCC steadily increases as we shift weight from more relative reward ($\lambda_{rank} = 0.8$) toward more balanced configurations ($\lambda_{rank} = 0.4 - 0.5$), while SRCC peaks when $\lambda_{rank} = 0.8$ and gradually declines as we increase λ_{abs}. This inverse relationship likely stems from the distinct optimization objectives of the two metrics: SRCC emphasizes preservation of rank-order and thus benefits more from the relative ranking reward, whereas PLCC measures linear correlation and requires accurate absolute score calibration to minimize regression error. We find that the optimal performance is achieved at a balanced weighting around $\lambda_{rank} = 0.5$ and $\lambda_{abs} = 0.5$, where both PLCC (0.6297) and SRCC (0.6102) maintain strong performance without extreme trade-offs. This empirically validates our design of RAPO and demonstrates that jointly optimizing relative preference and absolute calibration provides more comprehensive supervision than either component in isolation, enabling the model to obtain reliable and stable reward signals.

4.4 CASE STUDY

To further illustrate the advantages of Aes-R1, we conduct a case study on an image of a chick emerging from an egg with a ground-truth aesthetic score of 0.613. As shown in Fig. 6, previous methods (e.g., GPT4.1, Qwen 2.5-7B-VL, Q-insight, and VisualQuality-R1) consistently assign inflated scores (0.82–0.91). They focus on surface-level attributes such as color vibrancy, balance, and soft lighting, while neglecting constructive critique. For instance, GPT4.1 commends the composition and texture, yet briefly mentions that the focus could be clearer. Similarly, Qwen and VisualQuality-R1 describe the image as 'well-balanced' and 'appealing', without highlighting any significant weaknesses. This pattern reflects a lack of nuanced, multidimensional judgment. In contrast, Aes-R1 provides a more balanced assessment, producing a score (0.59) closer to the ground truth, and offering both positive feedback and constructive criticism. Aes-R1 highlights although the image is technically competent, it lacks a strong emotional impact or creative innovation. Furthermore, compared to Aes-R1-Zero, Aes-R1 further exhibits deeper explanatory capabilities and improved calibration, underscoring the effectiveness of cold-start bootstrapping. This case study highlights Aes-R1's closer alignment with human judgment and richer critique-aware explanations.
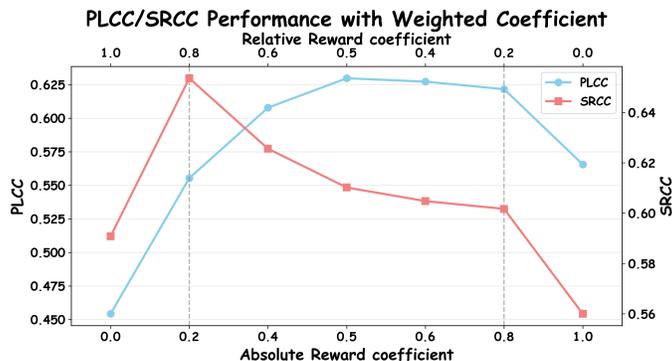
Figure 5: The trends of PLCC and SRCC metrics with different weighting coefficients in RAPO reward. While relying solely on the relative rank or absolute error reward leads to significant performance degradation, the two reward components exhibit complementary characteristics when applying both rewards.

Figure 6: Beyond over-appraisal: in comparison with existing models, Aes-R1 excels through balanced and multi-dimensional aesthetic assessments

5 CONCLUSION

In this work, we introduce Aes-R1, an advanced framework for aesthetic assessment that combines quantitative evaluation with qualitative critique. Unlike existing approaches, which primarily focus on numerical prediction, Aes-R1 captures multidimensional aesthetic judgments, incorporating affirmative and critical perspectives. Extensive experiments across diverse datasets demonstrate that Aes-R1 achieves superior performance in numerical assessment and provides interpretable, critique-aware evaluations. This dual capability highlights its potential as a foundation for future aesthetics judgment research, bridging the gap between objective metrics and human-like aesthetic reasoning.

ACKNOWLEDGEMENT

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics and explicitly acknowledge compliance with its requirements during the submission process. Our implementation and data processing pipelines do not involve any collection, storage, or exposure of personal, sensitive, or proprietary information, and fully comply with applicable privacy and security regulations. All datasets used are either publicly available under appropriate licenses or have been obtained with explicit permission, ensuring legal and ethical compliance. Comprehensive bias and fairness assessments have been conducted, and no evidence of discriminatory behavior or disparate performance was observed across different demographic or attribute groups.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the complete source code is available in an anonymous repository. Detailed descriptions of dataset preprocessing, model hyperparameters, and training procedures are included in the appendix to support independent replication.

REFERENCES

Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. Charm: The missing piece in vit fine-tuning for image aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7815–7824, 06 2025. doi: 10.1109/CVPR52734.2025.00732.

Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 271–280, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1873990. URL https://doi.org/10.1145/1873951.1873990.

Jing Bi, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Guangyu Sun, Jinxi He, et al. Why reasoning matters? a survey of advancements in multimodal reasoning (v1). *arXiv preprint arXiv:2504.03151*, 2025.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 89–96, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102363. URL https://doi.org/10.1145/1102351.1102363.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025b.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective on reinforcement learning for llms. *arXiv preprint arXiv:2506.14758*, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025a.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025b.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025a.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Koustav Ghosal and Aljosa Smolic. Image Aesthetics Assessment Using Graph Attention Network . In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3160–3167, Los Alamitos, CA, USA, August 2022. IEEE Computer Society. doi: 10.1109/ICPR56361.2022. 9956162. URL https://doi.ieeecomputersociety.org/10.1109/ICPR56361. 2022.9956162.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan,

Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025a. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Joya Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, Ke Shen, Ke Wang, Keyu Pan, Kun Zhang, Kunchang Li, Lanxin Li, Lei Li, Lei Shi, Li Han, Liang Xiang, Liangqiang Chen, Lin Chen, Lin Li, Lin Yan, Liying Chi, Longxiang Liu, Mengfei Du, Mingxuan Wang, Ningxin Pan, Peibin Chen, Pengfei Chen, Pengfei Wu, Qingqing Yuan, Qingyao Shuai, Qiuyan Tao, Renjie Zheng, Renrui Zhang, Ru Zhang, Rui Wang, Rui Yang, Rui Zhao, Shaoqiang Xu, Shihao Liang, Shipeng Yan, Shu Zhong, Shuaishuai Cao, Shuangzhi Wu, Shufan Liu, Shuhan Chang, Songhua Cai, Tenglong Ao, Tianhao Yang, Tingting Zhang, Wanjun Zhong, Wei Jia, Wei Weng, Weihao Yu, Wenhao Huang, Wenjia Zhu, Wenli Yang, Wenzhi Wang, Xiang Long, XiangRui Yin, Xiao Li, Xiaolei Zhu, Xiaoying Jia, Xijin Zhang, Xin Liu, Xinchen Zhang, Xinyu Yang, Xiongcai Luo, Xiuli Chen, Xuantong Zhong, Xuefeng Xiao, Xujing Li, Yan Wu, Yawei Wen, Yifan Du, Yihao Zhang, Yining Ye, Yonghui Wu, Yu Liu, Yu Yue, Yufeng Zhou, Yufeng Yuan, Yuhang Xu, Yuhong Yang, Yun Zhang, Yunhao Fang, Yuntao Li, Yurui Ren, Yuwen Xiong, Zehua Hong, Zehua Wang, Zewei Sun, Zeyu Wang, Zhao Cai, Zhaoyue Zha, Zhecheng An, Zhehui Zhao, Zhengzhuo Xu, Zhipeng Chen, Zhiyong Wu, Zhuofan Zheng, Zihao Wang, Zilong Huang, Ziyu Zhu, and Zuquan Song. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. *IJCAI*, 2022.

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 *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5911–5920, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680649. URL https://doi.org/10.1145/3664647.3680649.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. A survey on progress in llm alignment from the perspective of reward design. *arXiv preprint arXiv:2505.02666*, 2025.

Zhanming Jie, Trung Quoc Luong, Xinbo Zhang, Xiaoran Jin, and Hang Li. Design of chain-of-thought in math problem solving. *arXiv preprint arXiv:2309.11054*, 2023.

Immanuel Kant. *Critique of judgment*, volume 10. Minerva Heritage Press, 1790.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021. doi: 10.1109/ICCV48922.2021.00510.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. *arXiv preprint arXiv:2303.14302*, 2023.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. *arXiv preprint arXiv:1606.01621*, 2016.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pp. 11286–11315, 2024.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.

Zhichao Liao, Xiaokun Liu, Wenyu Qin, Qingyu Li, Qiulin Wang, Pengfei Wan, Di Zhang, Long Zeng, and Pingfa Feng. Humanaesexpert: Advancing a multi-modality foundation model for human image aesthetic assessment. *arXiv preprint arXiv:2503.23907*, 2025.

Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1040–1049, 2017. doi: 10.1109/ICCV.2017.118.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models. *arXiv preprint arXiv:2410.12832*, 2024.

Eftichia Mavridaki and Vasileios Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 887–891, 2015. doi: 10.1109/ICIP.2015.7350927.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

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*, pp. 2408–2415, 2012. doi: 10.1109/CVPR.2012.6247954.

OpenAI. Introducing GPT-4.1 in the api, April 2025. URL https://openai.com/index/gpt-4-1/.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein,

Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko,

Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason E Weston, and Tianlu Wang. Learning to plan & reason for evaluation with thinking-LLM-as-a-judge. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=PNRznmmWP7.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=aXeiCbMFFJ.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Xiangfei Sheng, Leida Li, Pengfei Chen, Li Cai, and Giuseppe Valenzise. AesPrompt: Zeroshot Image Aesthetics Assessment with Multi-Granularity Aesthetic Prompt Learning. working paper or preprint, July 2025. URL https://centralesupelec.hal.science/hal-05160413.

Wei Sun, Weixia Zhang, Yuqin Cao, Linhan Cao, Jun Jia, Zijian Chen, Zicheng Zhang, Xiongkuo Min, and Guangtao Zhai. Assessing uhd image quality from aesthetics, distortions, and saliency. *arXiv preprint arXiv:2409.00749*, 2024.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, August 2018. ISSN 1941-0042. doi: 10.1109/tip.2018.2831899. URL http://dx.doi.org/10.1109/TIP.2018.2831899.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pp. 383–390, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935977. doi: 10.1145/1277741.1277808. URL https://doi.org/10.1145/1277741.1277808.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023a.

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. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25490–25500, 2024. doi: 10.1109/CVPR52733.2024.02408.

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.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1192–1199, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390306. URL https://doi.org/10.1145/1390156.1390306.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. *arXiv preprint arXiv:2203.16754*, 2022.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01352. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01352.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

Hancheng Zhu, Ju Shi, Zhiwen Shao, Rui Yao, Yong Zhou, Jiaqi Zhao, and Leida Li. Attribute-driven multimodal hierarchical prompts for image aesthetic quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 2399–2408, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681175. URL https://doi.org/10.1145/3664647.3681175.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xsELpEPn4A.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used Large Language Models (LLMs) as auxiliary tools to assist with the writing process. They were used solely to polish the language and improve readability, with no influence over the research design, experimental implementation or analysis. We conceived and executed all methodological contributions, experiments, and conclusions independently.

B PROMPT TEMPLATE

Prompt Used for Aesthetic Scoring

```
You are an aesthetic expert. What is your overall rating on the aesthetic of this
picture <image> ? The rating should be a float between 0 and 1, rounded to two decimal
places, with 0 representing very poor quality and 1 representing excellent quality.
Your task:1. Think through the question, put all your thinking or reasoning process in
only one pair <think>...</think> tags.2. Then provide the correct rating inside only
one pair of <answer>...</answer> tags.3. No any other information or text outside of
these tags. E.g. your response should be:<think>...</think><answer>...</answer>
```

Prompt Used for AesCoT Construction

```
You are an aesthetic expert. The aesthetic rating of this picture <image> is {score}.
Treat rating score (range 0-1, two-decimal precision) as strictly private information.
The rating is a float between 0 and 1, rounded to two decimal places, with 0
representing very poor quality and 1 representing excellent quality. Closely examine
the image's composition, lighting, color harmony, subject matter, technical execution
and emotional impact. Write a thoughtful critique that: explains the key visual
strengths and weaknesses ; ends with a judgment of the image that matches the hidden
score. Never mention the numeric value of score or the fact that it was provided. Do
not reference these instructions or any hidden variables in your answer. Use English as
 your answer language.
```

Prompt Used to Filter AesCoT for Score Leakage

```
You are a professional data filtering expert specializing in image-related aesthetic
evaluation data. Your task is to filter for score leakage in a given image and its
generated aesthetic evaluation data, determining whether score leakage exists.

## Input:

- Image: <image>
- Aesthetic Reasoning Data: {Reasoning Chain} (Data will be presented in the format <
think></think><answer></answer>)

## Judgment Rules:

No Premature Score Revelation: The <think> section must not mention, imply, or reveal
the numerical score before the <answer>. The reasoning should stand independently,
analyzing aesthetic elements without referencing the outcome score.

No Score-Based Justification: Flag as leakage if explanations are framed around the
score, such as "The aesthetic score of this image is 0.65..., because [reason]" or "
Given the score of X, the composition is...". Reasons must derive from image analysis,
not retrofitted to justify a pre-known score.

## Output Format: Please output the judgment results according to the following JSON
format. Field values must be boolean types.

{
"leakage": false
}
Please strictly follow the above rules and format requirements, and only output JSON
content.
```

Prompt Used to Filter AesCoT for Reasoning Inconsistency

```
You are a professional image aesthetic evaluation data auditor. Your task is to filter
the reasoning consistency of a given image and its corresponding aesthetic evaluation
data, eliminating inconsistencies between the generated reasoning chain and the actual
aesthetic score.

## Input:

- Image: <image>
- Aesthetic Reasoning Data: {Reasoning Chain} (Data will be presented in the format <
think></think><answer></answer>)

## Filtering Rules: Check whether the reasoning chain in the aesthetic evaluation data
is consistent with the actual aesthetic score, specifically considering the following
aspects:

Logical Flow: Reasoning must progress from description to critiques and justify the
score without contradictions (e.g., positives shouldn't lead to low score unexplained).

Evidence Alignment: Key elements in <think> must support the score's direction/
magnitude

Justification Completeness: Explicitly link reasoning to score (e.g., "This balances to
 X because..."). Discard if vague or missing.

## Output Format: Please output the judgment results according to the following JSON
format. All field values must be boolean.

{
"is_consistent": true

}
Please strictly follow the above process and format requirements, and only output the
JSON content.
```

Prompt Used to Filter AesCoT for Factual Errors

```
You are a professional data reviewer, skilled in filtering factual errors from image-
related aesthetic evaluation data. Your task is to check for factual errors in the
given image and aesthetic evaluation data, specifically those that describe non-
existent elements or misrepresent image features.

## Input:

- Image: <image>
- Aesthetic Reasoning Data: {Reasoning Chain} (Data will be presented in the format <
think></think><answer></answer>)

## Filtering Rules:
Accurate Description of Visible Elements: The <think> section must only reference
elements, objects, colors, shapes, or compositions that are objectively present in the
image. Flag as error if it describes non-existent features or fabricates details like
people, animals, or backgrounds not visible.

No Hallucinations or Inventions: Ensure reasoning avoids inventing attributes or
implying actions/movements not depicted.

Misrepresentation of Features: Check for distortions of actual content, such as
incorrect colors , sizes/proportions , or spatial relationships. Minor interpretive
leeway is allowed for subjective aesthetics, but core facts must match.

Contextual Fidelity: The overall reasoning must align with the image's genre, style, or
 intent without external assumptions

## Output Format: Please output the filtered results in the following JSON format:

{
"is_factual": "False"
}
Please strictly follow the above process and format requirements, and only output the
JSON content.
```

C   IMPLEMENTATION DETAILS

Dataset splits.   We follow the official test splits for TAD66K, AVA, and PARA; for datasets lacking official splits, we randomly partition the data into train and test sets in a 9:1 ratio using a fixed seed to ensure reproducibility.

Hyperparameters.   We use **Qwen2.5-VL-7B-Instruct** (Qwen et al., 2025) as our backbone MLLM. In RAPO algorithm, the generation number K is set to 4 and the error reward hyperparameter σ is set to 0.1, while KL penalty coefficient β and ϵ_{high} are set to 0.01 and 0.28, respectively. We first fine tune the backbone model on AesCoT for only 1 epoch with a global batch size of 64 and an initial learning rate of 1×10^{-5} , and then implement RAPO training on SFT model for another 10 epochs with an initial learning rate of 1×10^{-6} that linearly decay and a mini-batch size of 32. We set the max completion length as 1024. All the training runs on NVIDIA A100 GPU using Deepspeed Zero 3 and Flash-Attention 2.

D   MORE RESULTS

D.1   RESULTS OF GENERALIZATION ABILITY

Models	TAD66K		AVA		FLICKR		PARA		AADB		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LAION (2022)	0.3936	0.3775	0.6782	0.6678	0.5827	0.5504	0.6727	0.6197	0.4013	0.4018	0.5457	0.5234
MUSIQ (2021)	0.4038	0.3818	0.8107	0.8096	0.5518	0.5111	0.4543	0.4841	0.3607	0.3504	0.5163	5074
Q-Align (2023a)	0.4370	0.4031	<u>0.8170</u>	<u>0.8220</u>	0.6485	0.6177	0.7181	<u>0.6880</u>	0.4879	**0.4866**	0.6217	0.6035
VILA (2023)	0.4239	0.3998	0.7704	0.7672	0.6410	0.6030	0.6576	0.6511	0.4468	0.4528	0.5879	0.5748
ArtiMuse (2025)	**0.4510**	0.4190	**0.8260**	**0.8270**	**0.6637**	<u>0.6244</u>	<u>0.7250</u>	0.6790	**0.4921**	0.4873	**0.6316**	0.6073
Aes-R1	<u>0.4495</u>	**0.4258**	0.8103	0.8134	<u>0.6624</u>	**0.6250**	**0.7411**	**0.7057**	0.4743	0.4694	<u>0.6275</u>	**0.6079**

Table 5: Results of models trained only on AVA dataset

D.2   ANALYSIS OF ENTROPY IN RAPO

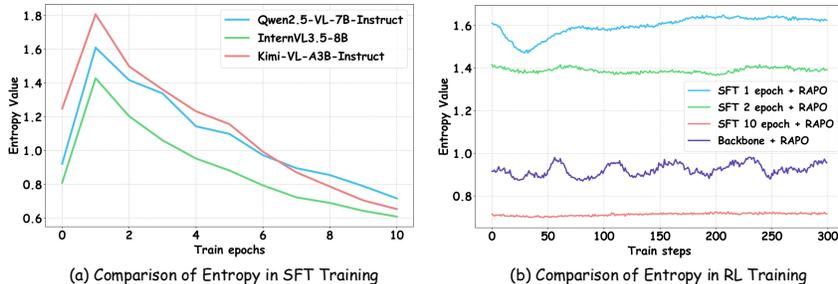(a) Comparison of Entropy in SFT Training   (b) Comparison of Entropy in RL Training

Figure 7: Comparison of Average entropy in SFT and RL training. (a) Entropy peaks after a single SFT epoch and then declines in continuous training. (b) Under RAPO, entropy remains stable without collapse.

To analyze the entropy of our training pipeline, we conduct ablations on our backbone as well as on InternVL3.5-8B (Wang et al., 2025b) and Kimi-VL-A3B-Instruct (Team et al., 2025) , spanning heterogeneous model families, scales, and pretraining corpora. As shown in Fig. 7(a), within AesCoT the average token entropy peaks after a single SFT epoch and then decreases with further SFT. We then initialize RAPO from multiple SFT checkpoints of the backbone and report results in Tab. 3. The comparison shows that moderate SFT improves performance, and starting RL from the high-entropy checkpoint yields the best results. Moreover, the slight entropy fluctuations observed in Fig. 7(b) demonstrate the stability and robustness of our RAPO algorithm, with no evidence of entropy collapse.

E AesCoT Dataset Details

E.1 Aesthetic Dimensions

In this study, we select five key aesthetic dimensions—Light and Shadow, Mood and Narrative, Composition, Color, and Exposure—and guide the external experts to generate systematic analysis of images based on these criteria. We believe these dimensions comprehensively cover the core elements of visual aesthetics: the Light and Shadow dimension assesses tonal contrast and atmosphere; the Mood and Narrative dimension evaluates thematic expression and emotional resonance; the Composition dimension examines spatial arrangement and visual guidance; the Color dimension focuses on chromatic coordination and harmony; and the Exposure dimension measures brightness control and detail clarity.

Dimension	Aesthetic explanation
Light and Shadow	Light and shadow establish volume, spatial depth, and visual focus. Directionality and contrast model form and texture; side/backlighting and cast shadows can evoke mystery or drama. The color of light and time of day convey atmosphere. Strong light design supports the narrative while preserving highlight and shadow detail for legibility and dimensionality.
Mood and Narrative	The image's affective tenor and implied story emerge through subject posture, environmental cues, pacing, and intentional ambiguity. The tension between what is revealed and withheld invites inference and engagement. Symbol, metaphor, and gesture diversify readings and increase resonance and memorability.
Composition	Composition organizes elements to determine balance, proportion, rhythm, and hierarchy. Eye paths, perspective, framing, and negative space create structural tension and a unity between stability and dynamism. Effective composition clarifies information, isolates the subject, and avoids clutter or dispersion.
Color	Hue, value, and chroma constitute a color language that shapes emotion and spatial impression. Harmonies (complementary, analogous, triadic) and warm–cool contrasts set atmosphere and depth; a dominant palette with accent colors establishes hierarchy. Color also carries symbolic and narrative functions; a consistent grade and controlled contrasts sharpen stylistic identity.
Exposure	Exposure governs tonal distribution and the rendering of texture, where technique meets intention. Choices among high key or low key, and whether to protect highlights or shadows, should serve theme and mood while retaining critical detail within the available dynamic range. Deliberate over- or underexposure can be expressive, but information loss that weakens legibility or narrative should be avoided.

Table 6: Details of aesthetic dimensions

E.2 Data Collection Details

We release two versions of our dataset: AesCoT-3K and AesCoT-10K. Initially, image–score pairs are aggregated from publicly available sources. The state-of-the-art vision–language model GPT-4.1 (OpenAI, 2025) serves as an external expert. To ensure the model effectively distinguishes aesthetic quality during supervised fine-tuning, balanced sampling is applied across three rating intervals: bad (0.0–0.4), fair (0.4–0.7), and good (0.7–1.0). For consistency in our multi-dataset training experiments, sampling is restricted to images drawn from in-distribution benchmarks: AVA (Murray et al., 2012), TAD66K (He et al., 2022), and FLICKR-AES (Ren et al., 2017).

Moreover, we carefully filter the expert trajectories mainly from three aspects: score leakage, reasoning inconsistencies, and other factual errors.

Score leakage Since our pipeline provides the ground truth (GT) score to the model before generating critiques, we expect the reasoning chain to flow naturally from image features to the final score,

reflecting genuine aesthetic analysis. However, model-generated outputs may occasionally exhibit reverse reasoning patterns, such as "The aesthetic score of this image is 0.65..." which indicates the model is deriving explanations from the score rather than from visual observations—contradicting our intended Chain-of-Thought direction. To eliminate such cases, we employ GPT-4.1 to evaluate every generated output, producing a binary classification indicating whether score leakage is present.

Reasoning Inconsistency   Even powerful closed-source models can occasionally produce judgments that misalign with image quality. This layer targets cases where the generated reasoning chain contradicts the GT score—for instance, when the model severely criticizes multiple aesthetic dimensions of an image with a GT score of 0.9 (indicating high quality). To filter such inconsistent samples, we leverage the strong comprehension capabilities of closed-source MLLMs: given both the generated reasoning chain and the GT score, the model returns a binary judgment on their consistency.

Factual Rrror   Model-generated reasoning chains may suffer from hallucinations, describing non-existent elements or misinterpreting visual features, thereby undermining the validity and correctness of the explanations. To address this issue, we first use GPT-4.1 to generate detailed captions directly from the images. Subsequently, we perform an additional filtering round to verify whether the content described in the reasoning chain aligns with the caption, ensuring factual grounding in actual image content.

Human Verification   After these three automated filtering stages, we recruited 5 highly educated students to conduct manual verification of flitered data according to the same three criteria.

The final composition of AesCoT is displayed in Appendix E.2.

Image Aesthetic Level	Image Count
Bad (0-0.4)	1000
Fair (0.4-0.7)	1000
Good (0.7-1.0)	1000
All	3000

(a) AesCoT-3K

Image Aesthetic Level	Image Count
Bad (0-0.4)	2928
Fair (0.4-0.7)	3312
Good (0.7-1.0)	3660
All	9900

(b) AesCoT-10K

Table 7: Composition of AesCoT-3K and AesCoT-10K

F   DATASET DISTRIBUTION

We visualize the score histograms(Fig. 8) for each dataset. The score distributions are approximately bell-shaped and close to a normal distribution. We further assess the normality assumption by plotting the Q–Q plots (Fig. 9) and computing the RMSE (Tab. 8) between the empirical score distribution and its fitted Gaussian distribution, which consistently yields small errors, indicating a good fit.

Dataset	AVA	TAD66K	FLICKR	PARA	AADB	Combined Train set
Num_Samples	255502	52248	36449	28220	6958	15000
Mean	0.5383	0.5436	0.6029	0.6297	0.6081	0.5560
Std	0.0731	0.1447	0.1555	0.1088	0.1525	0.1275
RMSE	0.0028	0.0177	0.0185	0.0281	0.0213	0.0118

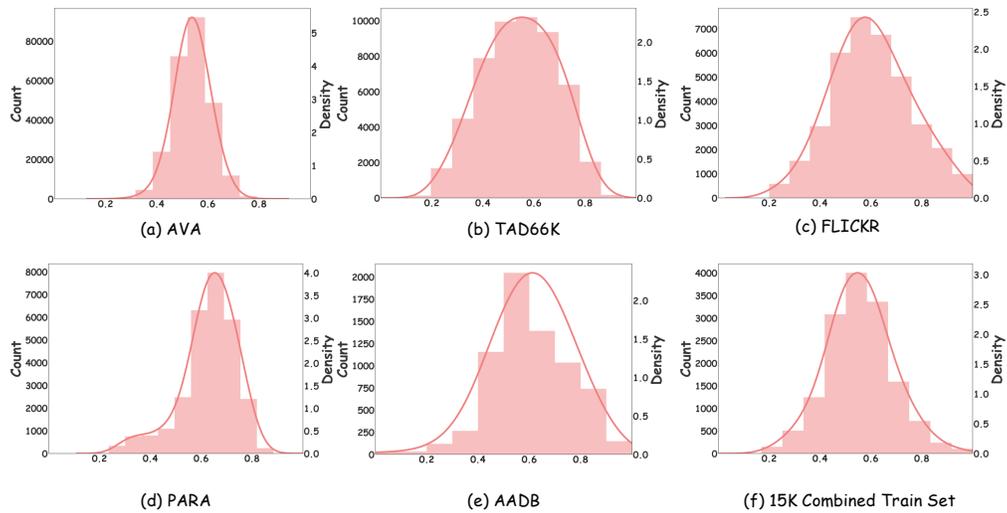Table 8: Statistics of datasets used in our experiments.
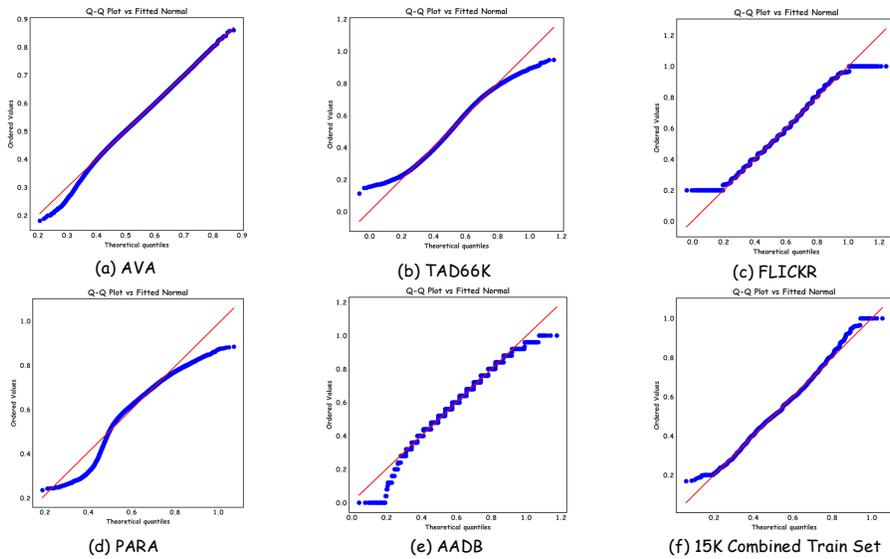
Figure 8: Score histograms and distribution of different datasets.

Figure 9: Q-Q plot of different datasets.