Beyond Absolute Scores: Relative Edit-induced Difference for Generalizable Image Aesthetic Assessment

Qifei Jia*[1], Xintong Yao*[1], Yasen Zhang†[1], Minghao Li[1], Yajie Chai[1], Qiming Lu[1], Baoyue Shen[1], Runyu Shi[1], Ying Huang[1], and Yue Zhang[2]

[1] Xiaomi Corporation, Beijing, China
{jiaqifei1, yaoxintong1, zhangyasen}@xiaomi.com
[2] Beijing University of Posts and Telecommunications, Beijing, China
zhangyuereal@bupt.edu.cn

Abstract. Traditional Image Aesthetic Assessment (IAA) methods mainly rely on regressing absolute Mean Opinion Scores (MOS). However, such a paradigm overlooks the inherently dynamic nature of human aesthetic perception, which relies on subconscious comparison against implicit visual references. Consequently, the lack of causal reasoning regarding aesthetic differences prevents models from learning generalizable aesthetic principles, thus limiting their generalization across diverse scenarios. In this work, we rethink the IAA task and propose **R**elative **E**dit-induced **D**ifference **Aes**thetic learning (**RED-Aes**), a novel framework that leverages controllable image editing models to simulate the human aesthetic reasoning process. Instead of fitting absolute score distributions, RED-Aes explicitly learns the visual factors that drive aesthetic changes. To support this paradigm, we construct the **RED-20k** dataset, which comprises editing-based image pairs, quantitative aesthetic differences, and Chain-of-Thought (CoT) reasoning. Furthermore, we introduce a three-stage training strategy guided by a relative ranking consistency reward, optimizing the model solely via relative supervision. Extensive experiments demonstrate that RED-Aes achieves state-of-the-art performance on multiple public benchmarks, exhibiting superior generalization capabilities.

Keywords: Image Aesthetic Assessment · Image Editing · Post Training

1 Introduction

Image Aesthetic Assessment (IAA) aims to quantify the aesthetic appeal of images. It underpins a growing range of real-world applications. In computational photography, on-device aesthetic models power best-shot selection and intelligent photo curation on modern smartphones [36]. Similarly, in the era of AI-generated content (AIGC), aesthetic scores serve as training-data filters [31] and

* Equal contribution.
† Corresponding author.

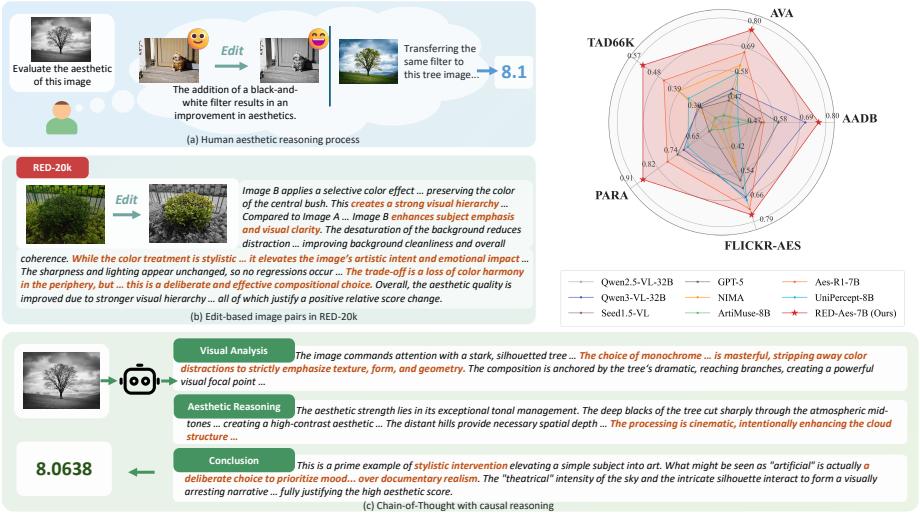

Fig. 1: Relative Difference Learning for Aesthetics. (a) Human aesthetic reasoning process: Illustration of how humans rely on implicit comparison to judge aesthetics. (b) Edit-based image pairs in RED-20k: An example of source-edit pairs where edit operations induce quantifiable aesthetic shifts. (c) CoT with causal reasoning: The model produces detailed textual explanations covering visual analysis and aesthetic reasoning to justify the predicted score.

reward signals for aligning text-to-image generative models [43]. Although recent works leverage Vision-Language Models (VLMs) [3, 25] for aesthetic assessment [24, 40, 41, 45], existing methods typically rely on crowdsourced Mean Opinion Scores (MOS) as supervision to regress absolute scores on limited datasets.

This paradigm overlooks a critical aspect of human aesthetic cognition. In practice, a human annotator rarely assigns an absolute score directly but instead relies on a reference anchor. This anchor is either an explicitly provided sample or an implicitly recalled image of similar content [22, 29] which allows the annotator to reason about differences to arrive at a score. While recent VLM-based approaches assess aesthetics from multiple dimensions [24, 41, 45], the causal link between reasoning and scores remains weak. Forcing models to fit score distributions without comparative knowledge constrains generalization. Human aesthetic judgment is thus a dynamic comparative process involving causal reasoning about aesthetic differences rather than an instantaneous absolute measurement, as illustrated in Fig. 1(a).

This insight has been validated in the Image Quality Assessment (IQA) domain where RankIQA [26] constructs quality ranking pairs by applying synthetic distortions of varying severity to source images to improve generalization. However, aesthetics is a high-level attribute dependent on content and semantics that cannot be manipulated through simple pixel-level operations. With the advancement of controllable image editing models [5, 46, 47], this idea can now be trans-

ferred to IAA. Targeted edits such as composition adjustments, color harmony enhancements, or lighting modifications create source-edit pairs (Fig. 1(b)). The aesthetic differences within these pairs encode explicit causal relationships where specific modifications directly induce quantifiable aesthetic shifts. This mechanism provides far stronger supervision than statistical correlations derived solely from absolute scores.

Building on this insight, we propose Relative Edit-induced Difference Aesthetic learning (RED-Aes), a framework that simulates how humans establish aesthetic concepts through comparison. RED-Aes explicitly learns relative aesthetic differences between source images and their edited counterparts to acquire generalizable principles that transfer across visual domains. To support this paradigm, we construct the RED-20k dataset through a fully automated data engine that leverages VLM reasoning models and controllable editing models to generate edited variants for each source image and obtain pairwise aesthetic difference scores with chain-of-thought (CoT) reasoning.

We design a three-stage training strategy. The first stage injects aesthetic difference knowledge via supervised fine-tuning (SFT) on source-edit pairs to endow the model with rich and attributable aesthetic representations. The second stage performs a lightweight score calibration. To fully leverage the aesthetic knowledge acquired from pre-training, the third stage employs Group Relative Policy Optimization (GRPO) [34] combined with a novel Relative Ranking Consistency Reward. This approach drives the model to produce predictions that are ordinally consistent within each source-edit group. This reinforcement learning stage enables the model to apply the causal aesthetic knowledge (Fig. 1(c)) acquired during pre-training more reliably, achieving superior generalization.

Our contributions are as follows:

- We construct RED-20k, an editing-based dataset of image pairs with quantitative aesthetic differences and CoT reasoning supported by a scalable data engine with multi-model ensemble and consensus-based quality control.
- We propose RED-Aes, a novel IAA framework that shifts from absolute score regression to relative edit-induced difference learning using a three-stage pipeline that progresses from comparative SFT to score calibration and GRPO-based reinforcement learning to optimize the model solely via relative supervision.
- Extensive experiments demonstrate state-of-the-art performance on five public benchmarks where our method outperforms both specialist and generalist models by significant margins.

2 Related Work

2.1 Image Aesthetic Assessment

Existing IAA datasets map single images to absolute aesthetic scores. AVA [28] provides approximately 250k crowdsourced ratings; TAD66K [13] offers theme-aware annotations; AADB [18] and PARA [44] enrich the space with multi-attribute labels and personalized preferences. On the modeling side, IAA has

evolved from hand-crafted features [8, 27] to deep learning. NIMA [36] popularized CNN-based MOS regression, followed by TANet [13], MUSIQ [17], and AesMamba [9] that improved architectures while retaining the same objective. More recently, VLMs have been applied to IAA. Q-Align [41] and Q-Instruct [40] leverage instruction tuning for visual scoring; AesExpert [15] generates expert-style descriptions; UNIAA [48] and ArtiMuse [7] combine scoring with fine-grained textual analysis. Despite these advances, the fundamental paradigm remains unchanged: mapping individual images to absolute scores via supervised training on limited datasets.

2.2 Controllable Image Editing

The rapid development of diffusion-based generative models has enabled precise controllable image editing, serving as the foundation of our data engine. InstructPix2Pix [5] pioneered instruction-guided editing by training a conditional diffusion model on synthetic paired data. Qwen-Image-Edit [39] is a 20B-parameter Multi-Modal Diffusion Transformer with outstanding semantic understanding capabilities. FLUX.1 Kontext [21] achieves high-fidelity editing through flow matching in latent space. LongCat-Image [37] is a 6B-parameter bilingual diffusion model trained with reward-model-guided reinforcement learning. Seedream [33] and Nano-Banana [10] are powerful closed-source models that deliver high-quality editing with robust performance. Using multiple architecturally distinct models ensures that aesthetic variations in our dataset reflect genuine visual changes rather than artifacts of any single pipeline.

2.3 Post-Training with Reinforcement Learning

DeepSeek-R1 [11] introduced a post-training paradigm combining supervised fine-tuning with reinforcement learning via GRPO [34], eliciting superior logical reasoning capabilities with significantly reduced computational costs. This paradigm has been rapidly adopted in VLM-based image assessment tasks. Q-Insight [23] applies GRPO for joint score regression and degradation perception. VisualQuality-R1 [42] trains an NR-IQA model with GRPO-based learning to rank under the Thurstone model. Aes-R1 [24] uses reinforcement learning to encourage more detailed aesthetic critiques. While these methods demonstrate the effectiveness of RL for assessment, they typically optimize for individual response quality or absolute score accuracy. In contrast, our approach employs GRPO with a Relative Ranking Consistency Reward that optimizes ordinal consistency within groups of aesthetically related images, directly enforcing comparative reasoning over causal aesthetic differences.

3 Methods

Our approach comprises two core components: (i) the RED-20k dataset, constructed through a fully automated pipeline leveraging VLMs and controllable

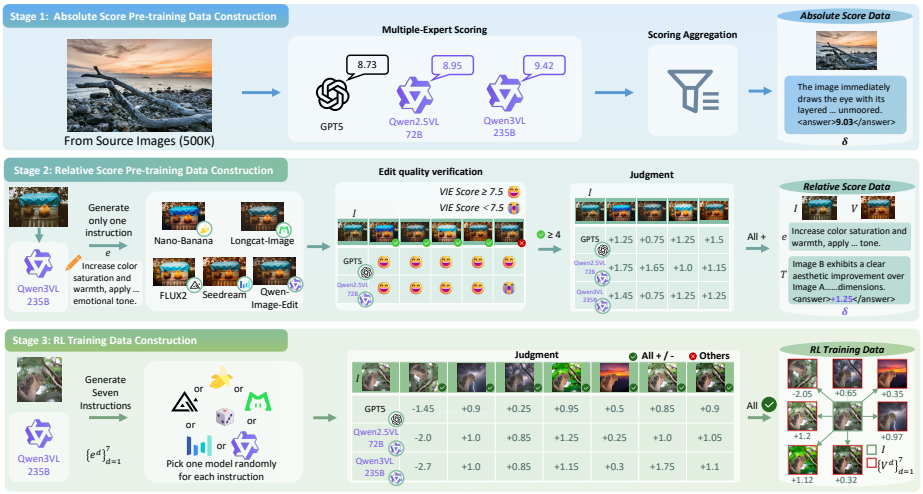

Fig. 2: Overview of the data construction pipeline. (1) Absolute Score Data Construction, utilizing multi-expert scoring combined with score aggregation strategies to obtain absolute scores; (2) Relative Score Data Construction, utilizing diverse editing models to generate source-edited pairs, filtering via VIEScores, and employing VLM judges for consistency verification; and (3) RL Training Data Construction, which scales up to group-wise samples by generating multiple instructions and applying random editors to create comprehensive preference datasets for reinforcement learning.

editing models to produce relative aesthetic score differences with causal reasoning (Sec. 3.1), and (ii) RED-Aes, a three-stage learning framework progressing from comparative learning to anchor calibration and GRPO-based optimization (Sec. 3.2). The detailed prompts used for VLMs throughout the data construction and training pipeline are provided in the supplementary material.

3.1 RED-20k: An Edit-based Aesthetic Difference Dataset

Source Data Collection. The dataset construction pipeline is illustrated in Fig. 2. We collect 500k images from properly licensed sources including Open-Images [19], DIV2K [1], UHD-IQA [14], Unsplash Lite [38], and proprietary in-house data, spanning diverse quality levels and content categories. Three VLMs (Qwen3-VL-235B, Qwen2.5-VL-72B, and GPT-5) independently score each image, and a multi-expert voting protocol produces a consensus score s_i for each image I_i, yielding a source pool $\mathcal{O} = \{(I_i, s_i)\}_{i=1}^{n}$, where n is the number of source images. These scores serve only as anchors for computing relative differences and generating editing instructions. The multi-expert scoring strategy utilizes MAD-based robust z-scores to detect outliers and adaptively selects a fusion method that employs weighted averaging under high consensus. Further details are provided in the supplementary material.

Pre-training Data. The construction proceeds in three steps. For each source image I with score s from $\mathcal{O}$, Qwen3-VL-235B first analyzes its aesthetic deficiencies and generates a targeted editing instruction e. We apply e to I using five editing models $\mathcal{M} = \{m_1, \ldots, m_5\}$ (Qwen-Image-Edit, FLUX2, Longcat-Image, Nano-Banana, and Seedream-4.5), producing five edited variants $\{V_j\}_{j=1}^{5}$, where $V_j = m_j(I, e)$.

Second, we employ the VIEScore [20] to verify edit quality. Two VLMs (Qwen2.5-VL-72B and GPT-5) compute VIEScores, and we apply a strict threshold of 7.5 on a [0,10] scale. A source image is retained only when at least four of five models produce successful edits:

$$\mathcal{V} = \{I,\, s,\, \{V_j^*\}_{j\in\mathcal{S}}\}, \quad |\mathcal{S}| \geq 4, \tag{1}$$

where $\mathcal{S} \subseteq \{1, \ldots, 5\}$ denotes the subset passing the quality threshold, V_j^* is the edited variant produced by model m_j that satisfies VIEScore ≥ 7.5, and $\mathcal{V}$ is the retained valid source-edit record.

Finally, three judgment models $\mathcal{J} = \{J_1, J_2, J_3\}$ (Qwen3-VL-235B, Qwen2.5-VL-72B, and GPT-5) independently assess every source-edit pair in $\mathcal{V}$, predicting the direction and magnitude of aesthetic change with CoT reasoning:

$$(\delta_{jk},\, T_{jk}) = J_k(I,\, V_j^*,\, s), \quad j \in \mathcal{S},\ k \in \{1, 2, 3\}, \tag{2}$$

where δ_{jk} denotes the aesthetic score difference (edited minus source) predicted by judge J_k for variant V_j^*, and T_{jk} is the corresponding CoT reasoning trace. A pair is retained only when all three judges unanimously agree on improved aesthetics. The final δ_j is computed via the voting protocol over $\{\delta_{jk}\}_{k=1}^{3}$. This strict consensus filters out noisy or ambiguous pairs. Among all valid variants, we randomly select one to form the final pre-training set $\mathcal{P}$, where each sample is a tuple $(I,\, s,\, V^*,\, \delta,\, T,\, e)$. Here V^* is uniformly sampled from $\{V_j^*\}_{j\in\mathcal{S}}$ and δ, T are the corresponding annotations.

RL Training Data. We uniformly sample 1k images from $\mathcal{O}$. For each source image I, Qwen3-VL-235B generates seven diverse editing instructions $\{e^d\}_{d=1}^{7}$, each sampled from a predefined pool of 16 edit categories (*e.g.*, style transfer, object removal, color grading). One editing model is randomly selected from $\mathcal{M}$ per instruction. The same multi-judge consensus is applied, retaining pairs only when all judges agree on the aesthetic change direction. The final RL group is:

$$\mathcal{G} = \{(I,\, s)\} \cup \left\{(V^d,\, s + \delta^d)\right\}_{d\in\mathcal{R}}, \tag{3}$$

where V^d is the edited variant, δ^d the consensus difference, and $\mathcal{R} \subseteq \{1, \ldots, 7\}$ the valid subset. Each group contains eight images (one source and seven variants) forming a comparison graph for GRPO.

Human Alignment Validation. We verify the reliability of labels produced by the VLMs. We conduct a human alignment study on 500 randomly sampled RED-20k pairs, with five evaluators including professional designers and lay users. The absolute aesthetic scores achieve an SRCC of 0.6843 against human annotations, while the relative source-edit differences, which serve as the core

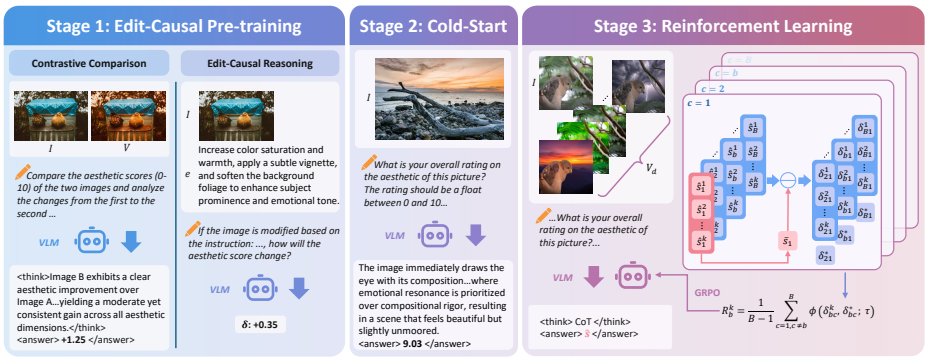

Fig. 3: The RED-Aes training pipeline. Stage 1 injects aesthetic knowledge through two SFT targets: PT-O1 compares source-edit pairs to predict aesthetic differences with CoT reasoning, while PT-O2 predicts the aesthetic impact of editing instructions without observing the edited image. Stage 2 aligns the model to output absolute aesthetic scores using only 200 anchor samples from the source pool. Stage 3 optimizes the policy through a relative ranking consistency reward that enforces ordinal agreement with ground-truth rankings.

supervision signal in our framework, reach a higher SRCC of 0.7379, confirming that relative comparison is more reliably aligned with human perception than absolute scoring.

3.2 RED-Aes: A Three-Stage Learning Framework

Stage 1: Edit-Causal Pre-training. The first stage injects aesthetic knowledge via supervised fine-tuning on RED-20k. As shown in Fig. 3, we focus on two optimization targets, contrastive comparison and edit-causal reasoning.

Contrastive Comparison (PT-O1). Given a source-edit pair (I, V), the model compares their aesthetic qualities, generates CoT reasoning describing the visual differences, and predicts the aesthetic difference δ.

Edit-Causal Reasoning (PT-O2). Given only the source image I and the editing instruction e (without the edited image), the model predicts the aesthetic difference δ that would result from the edit. This forces the model to internalize the causal relationship between visual modifications and aesthetic changes.

PT-O1 develops relative judgment from visual evidence, while PT-O2 extends this to causal prediction from textual instructions. Given the pre-training set $\mathcal{P}$, we minimize:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I,\, s,\, V^*,\, \delta,\, T,\, e)\sim\mathcal{P}} \left[\log p_\theta(\delta, T \mid \mathcal{X})\right], \tag{4}$$

where p_θ is the model distribution, $\mathcal{X} = (I, V^*)$ for PT-O1 and $\mathcal{X} = (I, e)$ for PT-O2. Both objectives are mixed and trained jointly for 3 epochs.

Stage 2: Cold-Start Calibration. After pre-training, the model possesses rich aesthetic knowledge through relative comparisons but cannot yet produce absolute scores. This stage bridges the gap via format alignment on 200 samples uniformly sampled from the source pool, each annotated with a standardized absolute score and CoT reasoning. It serves as a lightweight calibration preparing for reinforcement learning. We use the same hyperparameters as Stage 1 and train for 1 epoch.

Stage 3: Reinforcement Learning. The final stage employs GRPO [34] with a novel reward operating entirely on relative supervision (Fig. 3). Consider a batch of B images from the same group $\mathcal{G}$. For each image I_b, the policy π_θ samples K outputs with predicted score $\hat{s}_b^k$. Let δ_{bc}^* denote the ground-truth score difference between I_b and I_c. Each image serves as a comparison anchor. We compute the mean prediction $\bar{s}_c = (1/K)\sum_{k'=1}^{K} \hat{s}_c^{k'}$ for every other image I_c and define the relative-ranking reward R_b^k for the k-th output of image I_b as:

$$R_b^k = \frac{1}{B-1} \sum_{c=1, c\neq b}^{B} \phi\big(\hat{s}_b^k - \bar{s}_c,\ \delta_{bc}^*;\ \tau\big)\,, \tag{5}$$

where $\phi(\delta_{\text{pred}}, \delta_{\text{gt}}; \tau) = \tanh(|\delta_{\text{gt}}|) \cdot \sigma(\delta_{\text{pred}} \cdot \delta_{\text{gt}}/\tau)$ is a soft Kendall concordance function, $\sigma(\cdot)$ the sigmoid, and τ a temperature. The tanh term serves as a gap-aware weight emphasizing pairs with large aesthetic differences while suppressing near-ties, and the sigmoid measures soft agreement between predicted and ground-truth ordinal directions.

The advantage is computed relative to group statistics:

$$A_b^k = \frac{R_b^k - \text{mean}(\{R_b^{k'}\}_{k'=1}^{K})}{\text{std}(\{R_b^{k'}\}_{k'=1}^{K})}, \tag{6}$$

and the policy is updated by maximizing the following GRPO objective:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{q\sim\mathcal{D},\ \{o_b^k\}_{k=1}^{K}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{K} \sum_{k=1}^{K} \min\big(r_b^k A_b^k,\ \hat{r}_b^k A_b^k\big) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \tag{7}$$

where $\mathcal{D}$ is the RL training distribution, q is the input prompt, o_b^k is the k-th sampled response for image I_b, $\pi_{\theta_{\text{old}}}$ is the old policy used for importance sampling, $\hat{r}_b^k = \text{clip}(r_b^k, 1-\epsilon, 1+\epsilon)$, $r_b^k = \pi_\theta(o_b^k \mid q) \,/\, \pi_{\theta_{\text{old}}}(o_b^k \mid q)$ is the importance sampling ratio, ϵ is the clipping range, and β controls the strength of the KL divergence penalty against the reference policy π_{ref}. This optimization encourages the model to produce scores that are ordinally consistent within each source-edit group, without requiring any absolute score supervision during the RL stage. In our implementation, we use $B = 8$, $K = 6$, and $\tau = 0.1$.

4 Experiments

We begin by demonstrating the zero-shot generalization of our model on five benchmarks where it outperforms both supervised aesthetic methods and gen-

Table 1: Computational cost of RED-20k construction.

Stage	GPU Hours (H200)	API Calls
Absolute Score Data Construction	695	500k
Relative Score Data Construction	1,100	500k
Pairwise Aesthetic Judging	410	295k
RL Training Data Construction	15	1.5k
7B SFT Training	12	—
7B RL Training	560	—

eralist VLMs without accessing target training data. We subsequently conduct ablation studies to validate that source-consistent comparative learning effectively decouples aesthetic assessment from semantic biases. We also verify the synergy between content-oriented and quality-oriented editing instructions while determining the optimal ensemble trade-off for our data engine. We further confirm that our edit-causal learning objective and Relative Consistency Reward substantially surpass standard contrastive objectives and reasoning-based baselines. We finally investigate the impact of reinforcement learning group size and show that a contrastive context involving one source image against seven edited variants is essential for learning fine-grained rankings.

4.1 Experiment Settings

Datasets and Metrics. We evaluate our method on five public benchmarks including TAD66K [13], AVA [28], FLICKR-AES [30], PARA [44], and AADB [18]. All experiments operate under a cross-domain setting where the model is trained exclusively on our proposed dataset. We normalize the ground-truth Mean Opinion Scores from their original ranges of 1-10 for TAD66K and AVA or 1-5 for others to the [0, 1] interval. We utilize Pearson Linear Correlation Coefficient (PLCC) [4] to measure prediction accuracy and Spearman Rank-Order Correlation Coefficient (SRCC) [32] to assess monotonicity.

Evaluation Settings. We report results on all five datasets for main comparisons and conduct ablation studies on the larger TAD66K, AVA, and FLICKR-AES benchmarks to ensure statistical robustness. All models function in a zero-shot manner without fine-tuning on target datasets. We run open-source models using recommended inference settings for baselines that lack public results. The prompts used for all evaluated models are provided in the supplementary material to ensure reproducibility.

Implementation Details. We implement the RED-Aes using Qwen2.5-VL-7B [3], Qwen2.5-VL-3B [3], and Qwen3-VL-2B [2] backbones with frozen vision encoders and a fine-tuned language component. The training pipeline utilizes the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ plus weight decay 0.01 and resizes images to 448×448. We first conduct supervised fine-tuning on our synthesized dataset for 3 epochs with a batch size of 16 and a learning rate of

Table 2: Quantitative evaluation on multiple public benchmarks. † Results are reproduced by us using the official open-source models as the original papers did not report performance on these specific datasets.

Model	TAD66K		AVA		FLICKR		PARA		AADB		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Vanilla VLM												
Qwen2.5-VL-7B [3]	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 [3]	0.2300	0.2715	0.3893	0.4170	0.5800	0.6527	0.7176	0.7118	0.5060	0.5138	0.4846	0.5134
Qwen3-VL-8B [2]	0.2642	0.2464	0.4663	0.4599	0.6258	0.6190	0.6050	0.6127	0.6934	0.6460	0.5309	0.5168
Qwen3-VL-32B [2]	0.3042	0.2816	0.5057	0.4955	0.6651	0.6545	0.7176	0.7118	0.7147	0.6782	0.5815	0.5643
Seed1.5-VL-thinking [12]	0.2908	0.2974	0.4533	0.4871	0.5837	0.6374	0.6134	0.6364	0.5279	0.5577	0.4938	0.5232
GPT-4o [16]	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-5 [35]	0.2973	0.3347	0.4852	0.5447	0.6218	0.6260	0.7433	0.7352	0.6001	0.5989	0.5495	0.5679
Multi-Expert Voting	0.4925	0.5031	0.6654	0.6743	0.7175	0.7219	0.8397	0.7957	0.7124	0.6765	0.6855	0.6743
Deep-Learning Based												
NIMA [36]	0.3885	0.3654	0.6120	0.6361	0.5130	0.4796	0.5868	0.5709	0.3886	0.3904	0.4978	0.4885
VLM Based												
ArtiMuse-8B [7]	0.2320	0.2300	0.3850	0.3970	0.3340	0.3490	0.6057†	0.5624†	0.4836†	0.5130†	0.4081	0.4103
Aes-R1-7B [24]	0.4513	0.4248	0.6702	0.6619	0.7243	0.6973	0.7842	0.7666	0.5386	0.5423	0.6337	0.6186
UniPercept-8B [6]	0.3460	0.3360	0.5770	0.5890	0.6810	0.6880	0.7006†	0.5869†	0.4285†	0.4237†	0.5466	0.5247
RED-Aes-7B (Ours)	**0.5406**	**0.5322**	**0.7713**	**0.7700**	0.7524	0.7252	**0.8846**	**0.8591**	**0.7722**	**0.7746**	**0.7442**	**0.7322**
RED-Aes-3B (Ours)	0.5275	0.5297	0.7518	0.7514	0.7292	0.7089	0.8547	0.8413	0.7365	0.7509	0.7199	0.7164
RED-Aes-2B (Ours)	0.4784	0.4778	0.7235	0.7148	**0.7565**	**0.7401**	0.8616	0.8360	0.7426	0.7523	0.7125	0.7042

1e-5. We subsequently execute a format alignment stage for 1 epoch using 200 samples to standardize scalar output. We finally employ reinforcement learning via GRPO to optimize ranking consistency by constructing training groups of 8 images and generating 6 sampled outputs per prompt.

Computational Cost. During dataset construction, the open-source Qwen-series models are deployed on H200 GPUs for local inference, while GPT-5 is accessed via API calls. These large-scale models are used exclusively for offline dataset construction. Tab. 1 details the computational cost of each construction stage and model training, including GPU hours for locally deployed models and the number of API calls. Furthermore, the RL stage operates on only 1k source images and their corresponding edited variants, and the entire annotation and editing pipeline is sample-independent and highly parallelizable.

4.2 Comparison with State-of-the-Art

We evaluate our method on five benchmarks including TAD66K, AVA, FLICKR, PARA, and AADB under a rigorous zero-shot cross-dataset setting. Our model is trained exclusively on the proposed dataset. As shown in Tab. 2, our model based on Qwen2.5-VL-7B achieves state-of-the-art performance with an average PLCC of 0.7442 and SRCC of 0.7322. It substantially surpasses UniPercept [6] and GPT-5, which represent the leading generalist aesthetic models. Furthermore, it outperforms Aes-R1 which serves as the strongest specialist model trained directly on in-domain data. This performance is particularly notable considering that our approach generalizes from synthesized comparative pairs, whereas baselines rely heavily on fitting target-domain absolute scores. Remarkably, even our lightweight variant based on Qwen3-VL-2B outperforms all baselines to demonstrate exceptional efficiency. We also compare against a zero-shot multi-expert voting baseline that directly ensembles scores from the three VLMs used in our

Table 3: Ablation Study on Dataset Quality and Training Consistency. We progressively evaluate the contribution of our collected dataset (Rows 1-2) and the impact of the group-wise consistent training strategy (Rows 3-4) across three benchmarks.

Configuration	AVA		FLICKR		TAD66K		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Aes-R1 *w/* AesCoT-15k	0.6702	0.6619	0.7243	0.6973	0.4513	0.4248	0.6153	0.5947
Aes-R1 *w/* RED-20k	0.7506	0.7381	0.7270	0.6983	0.5107	0.5032	0.6628	0.6465
Ours (Shuffled Batch)	0.7564	0.7423	0.7281	0.7038	0.5184	0.5134	0.6676	0.6532
Ours (Consistent Batch)	**0.7713**	**0.7700**	**0.7524**	**0.7252**	**0.5406**	**0.5322**	**0.6881**	**0.6758**

data construction (Qwen3-VL-235B, Qwen2.5-VL-72B, and GPT-5). This ensemble achieves an average SRCC of only 0.6743 with prohibitive deployment cost, falling below even our 2B model, confirming the effectiveness of our training paradigm in distilling aesthetic reasoning capabilities into compact models.

We attribute these gains to the comparative supervision provided by our rigorously constructed dataset. First, the relative-score pre-training stage injects comparative reasoning capabilities that allow the model to learn universal aesthetic gradients rather than overfitting to absolute scalar values. Second, the reinforcement learning stage utilizes our relative ranking consistency reward to enforce ordinal relationships among images. This design aligns with the inherently comparative nature of human perception. The consistent improvements observed across Qwen2.5-VL and Qwen3-VL backbones further validate that our method is architecture-agnostic and generalizes effectively across different VLM families.

4.3 Efficacy of Dataset and Group-wise Consistency

To disentangle the contributions of our dataset from our training paradigm, we conduct a progressive ablation study as presented in Tab. 3.

Dataset Superiority. We first validate the quality of our collected dataset by replicating the training pipeline of the Aes-R1 baseline but substituting their original data with our proposed dataset. As indicated in the comparison between the first two rows, merely transitioning to our dataset yields a substantial performance gain where the average SRCC increases from 0.5947 to 0.6465. This significant margin confirms that our dataset provides supervision signals of superior quality and robustness compared to prior benchmarks.

Importance of Consistent Context. Next, we investigate the impact of group-wise consistency. We compare our strategy against a baseline that shuffles images within the batch while preserving the pairing between each image and its original editing instruction, thereby forcing comparisons between unrelated content. Results demonstrate that the shuffled setting underperforms our consistent approach with an average SRCC of 0.6532 versus 0.6758. We attribute this gap to content-dependent biases, as comparing disparate images leads the model to

Table 4: Ablation study on editing instruction types. We report PLCC and SRCC across three benchmarks. "Cont." and "Qual." denote content-oriented and quality-oriented editing instructions, respectively. "PT" refers to Pre-training and "RL" to Reinforcement Learning. "Mixed" implies using both instruction types.

Method Setting	AVA		FLICKR		TAD66K		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Cont. Only PT & RL	0.7050	0.6939	0.7234	0.6987	0.5144	0.5020	0.6476	0.6315
Qual. Only PT & RL	0.7152	0.7130	0.7221	0.6944	0.5238	0.5188	0.6537	0.6421
Mixed PT + Qual. RL	0.7426	0.7379	0.7450	0.7047	0.5317	0.5202	0.6731	0.6543
Mixed PT & RL	**0.7713**	**0.7700**	**0.7524**	**0.7252**	**0.5406**	**0.5322**	**0.6881**	**0.6758**

learn spurious correlations such as category preferences instead of genuine aesthetic improvements. By contrast, our strategy isolates aesthetic changes from the content so that the model focuses purely on the relative quality of edits.

4.4 Impact of Edit Types

We categorize editing instructions into content-oriented tasks (e.g., inpainting) and quality-oriented tasks (e.g., color enhancements). Tab. 4 presents the evaluation of models trained on various combinations of these types.

Synergy between Content and Quality. We observe that neither content-only nor quality-only training yields optimal performance. The full model integrates both types during pre-training and reinforcement learning stages and consistently outperforms the ablated variants. For instance, it increases the average SRCC from 0.6421 in the quality-only setting to 0.6758. This confirms that robust assessment benefits from understanding both low-level perceptual attributes and high-level semantic composition.

Necessity of Diverse RL Supervision. We further compare the full model against the variant using mixed pre-training with quality-only reinforcement learning. Excluding content edits from the alignment stage leads to a distinct performance drop across all datasets. Specifically, the SRCC on AVA falls from 0.7700 to 0.7379. These results demonstrate that reinforcing ranking consistency across diverse edit types is crucial for generalization.

4.5 Impact of Ensemble Size

Number of Editing Models (N_E). We conducted a pilot study on 5k raw images to determine the optimal configuration. We first investigated editing diversity by varying N_E from 1 to 7. Tab. 5 shows that increasing N_E to 5 yields substantial PLCC and SRCC gains and effectively captures diverse aesthetic causality. A further increase to 7 offers only marginal improvements. This indicates that five editing models sufficiently cover the variation space.

Number of Judgment Models (N_J). We subsequently analyzed the consensus mechanism with N_E fixed at 5. Transitioning from 1 to 3 judgment models

Table 5: Ablation study on the ensemble size of editing and judgment models. We denote N_E as the number of editing models and N_J as the number of aesthetic judgment models. The chosen setting (5+3) achieves the best trade-off between performance and computational cost.

Number Setting		AVA		FLICKR		TAD66K		Average	
N_E	N_J	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
1	1	0.6940	0.6834	0.6327	0.6106	0.4917	0.4876	0.6061	0.5939
3	1	0.7379	0.7165	0.6800	0.6477	0.5159	0.4969	0.6446	0.6204
5	1	0.7442	0.7309	0.7055	0.6667	0.5224	0.5029	0.6574	0.6335
7	1	0.7459	0.7346	0.7089	0.6630	0.5214	0.5073	0.6587	0.6350
5	3	0.7713	0.7700	0.7524	**0.7252**	**0.5406**	0.5322	0.6881	0.6758
5	5	**0.7735**	**0.7722**	**0.7531**	0.7234	0.5383	**0.5330**	**0.6883**	**0.6762**

significantly enhances performance by filtering label noise. Increasing N_J to 5 yields negligible average gains and even slight degradation on generalization benchmarks. The configuration with 5 editing and 5 judgment models imposes a higher computational burden without meaningful benefits. We therefore adopted the strategy of 5 editing and 3 judgment models as the optimal trade-off.

4.6 Component Analysis

Effectiveness of Pre-training. We first examine pre-training objectives by training the model on the full dataset with supervised fine-tuning (SFT). Our dataset provides absolute aesthetic scores for all images, derived from expert ratings and relative comparisons, enabling standard SFT on the entire data. As shown in Phase 1 of Tab. 6, the SFT only baseline achieves an average SRCC of only 0.5577. Adding our contrastive objective (PT-O1) raises SRCC to 0.5809. Further incorporating edit causal learning (PT-O2) boosts it to 0.6218. This improvement validates that combining static representation learning with dynamic causal modeling outperforms vanilla SFT.

Superiority of Our RL. In Phase 2, we evaluate RL strategies starting from a cold-start base. This base is constructed by first pre-training the model on the full dataset with PT-O1 & PT-O2 to learn comparative aesthetic knowledge, followed by a cold-start supervised fine-tuning (SFT) on only 200 samples with absolute score annotations to enable calibrated absolute score output (see Sec. 4.1). Due to limited data, its SRCC drops to 0.5907, which is expected and provides a controlled starting point for RL comparison. On this base, Aes-R1 improves SRCC to 0.6465. Our RL method further pushes it to 0.6758, surpassing both Aes-R1 and the fully-supervised Phase 1 model (0.6218). This demonstrates that our relative-aesthetic reward aligns better with human preferences than general-purpose reasoning RL.

Table 6: Component ablation study. We evaluate the incremental effectiveness of pre-training objectives (Phase 1) and reinforcement learning strategies (Phase 2) on three aesthetic benchmarks.

Method Setting	AVA		FLICKR		TAD66K		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Phase 1: Pre-training Strategies (Full SFT)								
SFT only	0.6612	0.6332	0.6618	0.6390	0.4344	0.4009	0.5858	0.5577
SFT *w/* PT-O1	0.6849	0.6685	0.6878	0.6456	0.4550	0.4287	0.6092	0.5809
SFT *w/* PT-O1 & PT-O2	0.7398	0.7129	0.7159	0.6733	0.4906	0.4791	0.6488	0.6218
Phase 2: RL Strategies (Few-shot Cold Start)								
Cold Start Base	0.7023	0.6790	0.6911	0.6510	0.4719	0.4421	0.6218	0.5907
Cold Start *w/* Aes-R1	0.7506	0.7381	0.7270	0.6983	0.5107	0.5032	0.6628	0.6465
Cold Start *w/* Our RL	**0.7713**	**0.7700**	**0.7524**	**0.7252**	**0.5406**	**0.5322**	**0.6881**	**0.6758**

Table 7: Ablation on the Group Size (Batch Size) in RL. In our framework, a "batch" consists of 1 source image and $B-1$ edited variants. We study how the number of edited candidates per image affects the optimization. $B = 8$ (1 source + 7 edits) provides the most effective contrastive context.

Batch Size (B)	AVA		FLICKR		TAD66K		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
4 (1+3)	0.7579	0.7483	0.7325	0.7110	0.5205	0.5160	0.6703	0.6584
6 (1+5)	0.7676	0.7599	0.7434	0.7240	0.5347	0.5246	0.6819	0.6695
8 (1+7)	**0.7713**	**0.7700**	0.7524	0.7252	**0.5406**	**0.5322**	**0.6881**	**0.6758**
10 (1+9)	0.7701	0.7688	**0.7576**	**0.7257**	0.5361	0.5267	0.6879	0.6737

4.7 Impact of RL Group Size

Effect of Contrastive Diversity. We investigate the impact of the group size B (comprising one source image and $B-1$ edited variants) on the optimization process. As shown in Tab. 7, a small group size ($B = 4$) yields suboptimal performance, achieving an average SRCC of only 0.6584, because the model observes limited aesthetic variations, restricting its ability to distinguish subtle improvements from noise. Increasing B to 8 brings significant gains, boosting the average SRCC to 0.6758. This richer contrastive context enables the model to compare the source against a diverse set of superior and inferior variants, facilitating the learning of fine-grained ranking signals. However, further expanding the size to $B = 10$ leads to performance saturation and a slight decline in the average SRCC. This suggests that 7 edited variants are sufficient to estimate the local aesthetic gradient, while adding excessive candidates may introduce redundancy or increase optimization difficulty. We therefore adopt $B = 8$ as the default setting.

Fig. 4: Qualitative comparison of RED-Aes against baselines. Each column displays the input images with Ground Truth score, and the predicted scores from each model, along with the CoT reasoning.

4.8 Case Study

As illustrated in Fig. 4, both generic VLMs and absolute-score baselines struggle to decouple semantic content from aesthetic quality. In the second row, GPT-5 and Aes-R1 assign inflated scores of 8.22 and 3.70 to a heavily blurred image (GT 1.72), failing to penalize degradation. RED-Aes predicts 3.11 by explicitly critiquing the "excessive soft focus" and "synthetic effect". In the fourth row, despite severe artifacts (GT 2.17), Aes-R1 and GPT-5 still predict 8.05 and 8.75, misled by semantic content such as the dog's "expressive face". RED-Aes correctly identifies the "garish saturation" and "muddled texture," yielding 3.17. These results confirm that our relative edit-induced learning effectively diagnoses the specific visual factors that influence aesthetic scores.

5 Conclusion

We propose RED-Aes, a framework that shifts IAA from absolute score regression to relative edit-induced difference learning. By constructing source-edit pairs via controllable editing models, our approach enables the model to learn causal factors underlying aesthetic perception. The accompanying RED-20k dataset provides the first resource of editing-based aesthetic pairs with quantitative differences and CoT reasoning, built through a fully automated pipeline without human annotation. To progressively build aesthetic reasoning capabilities, we employ a three-stage training strategy comprising edit-causal pre-training, format alignment, and GRPO with a relative ranking consistency reward. Experiments on five benchmarks under zero-shot cross-domain evaluation demonstrate state-of-the-art performance, with even a lightweight 2B variant surpassing all existing baselines.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025), `https://api.semanticscholar.org/CorpusID:276449796`
4. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cao, S., Li, J., Li, X., Pu, Y., Zhu, K., Gao, Y., Luo, S., Xin, Y., Qin, Q., Zhou, Y., et al.: Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. arXiv preprint arXiv:2512.21675 (2025)
7. Cao, S., Ma, N., Li, J., Li, X., Shao, L., Zhu, K., Zhou, Y., Pu, Y., Wu, J., Wang, J., et al.: Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. arXiv preprint arXiv:2507.14533 (2025)
8. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: European conference on computer vision. pp. 288–301. Springer (2006)
9. Gao, F., Lin, Y., Shi, J., Qiao, M., Wang, N.: Aesmamba: Universal image aesthetic assessment with state space models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7444–7453 (2024)
10. Google: Nano Banana Pro: Gemini 3 Pro Image model from Google DeepMind. `https://blog.google/innovation-and-ai/products/nano-banana-pro/` (11 2025), accessed: 2026-03-19
11. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
13. He, S., Zhang, Y., Xie, R., Jiang, D., Ming, A.: Rethinking image aesthetics assessment: Models, datasets and benchmarks. In: IJCAI. vol. 6, p. 22 (2022)
14. Hosu, V., Conde, M.V., Agnolucci, L., Barman, N., Zadtootaghaj, S., Timofte, R., Sun, W., Zhang, W., Cao, Y., Cao, L., et al.: Aim 2024 challenge on uhd blind photo quality assessment. In: European Conference on Computer Vision. pp. 261–286. Springer (2024)
15. Huang, Y., Sheng, X., Yang, Z., Yuan, Q., Duan, Z., Chen, P., Li, L., Lin, W., Shi, G.: Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5911–5920 (2024)

16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
18. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European conference on computer vision. pp. 662–679. Springer (2016)
19. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages **2**(3), 18 (2017)
20. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12268–12290 (2024)
21. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
22. Leder, H., Belke, B., Oeberst, A., Augustin, D.: A model of aesthetic appreciation and aesthetic judgments. British journal of psychology **95**(4), 489–508 (2004)
23. Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., Zhang, J.: Q-insight: Understanding image quality via visual reinforcement learning. arXiv preprint arXiv:2503.22679 (2025)
24. Liu, B., Hu, Y., Jin, S., Dou, S., Shi, G., Shao, J., Gui, T., Huang, X.: Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. arXiv preprint arXiv:2509.21871 (2025)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
26. Liu, X., van de Weijer, J., Bagdanov, A.D.: Rankiqa: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
27. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 international conference on computer vision. pp. 1784–1791. IEEE (2011)
28. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2408–2415. IEEE (2012)
29. Reber, R., Schwarz, N., Winkielman, P.: Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality and social psychology review **8**(4), 364–382 (2004)
30. Ren, J., Shen, X., Lin, Z., Mech, R., Foran, D.J.: Personalized image aesthetics. In: Proceedings of the IEEE international conference on computer vision. pp. 638–647 (2017)
31. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)

32. Sedgwick, P.: Spearman's rank correlation coefficient. Bmj **349** (2014)
33. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
34. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
35. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
36. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018)
37. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., et al.: Longcat-image technical report. arXiv preprint arXiv:2512.07584 (2025)
38. Unsplash: Unsplash lite dataset 1.3.0. `https://unsplash.com/data` (2020)
39. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
40. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., et al.: Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25490–25500 (2024)
41. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
42. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. arXiv preprint arXiv:2505.14460 (2025)
43. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
44. Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., Guo, Y.: Personalized image aesthetics assessment with rich attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19861–19869 (2022)
45. You, Z., Gu, J., Li, Z., Cai, X., Zhu, K., Dong, C., Xue, T.: Descriptive image quality assessment in the wild. arXiv e-prints pp. arXiv–2405 (2024)
46. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
47. Zhao, H., Ma, X., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)
48. Zhou, Z., Wang, Q., Lin, B., Su, Y., Chen, R., Tao, X., Zheng, A., Yuan, L., Wan, P., Zhang, D.: Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. arXiv preprint arXiv:2404.09619 (2024)