From Mimesis to Metamorphosis: Evolving VLM Judges via In-Context Comparing and Knowledge Internalization

Juntuo Wang[1,2]*, Yuming Qiao[1]*, Yifan Yang[1]*, Lunxi Yuan[1], Liang Luo[1]†, Dan Meng[1]†

[1]OPPO AI Center, [2]Brown University

Correspondence: luoliang1@oppo.com, mengdan90@163.com

Abstract

Vision-language models (VLMs) are increasingly adopted as judges for subjective assessment, yet absolute scoring remains brittle due to inconsistent scales and inherent preference biases. To bridge this gap, we propose S^{2}AD **(Semantic-Anchored Scale-Agnostic Distillation)**, a novel easy-to-hard framework that operationalizes subjective assessment as comparative analysis, conceptualizing the judge's evolution from mimesis to metamorphosis. In Stage 1 (Mimesis), we introduce Dynamic Soft Positioning (DSP) to train the judge to compare a query against retrieved reference images, establishing a relative evaluation space that ensures consistent ordering under heterogeneous scales. In Stage 2 (Metamorphosis), this comparative capability is internalized via Language Buttons—discrete semantic levels serving as a retrieval-free internal reference. Optimized with Group Relative Policy Optimization (GRPO), S^{2}AD achieves efficient, scale-steerable inference that adapts to diverse grading standards. Our framework reaches state-of-the-art performance across multiple benchmarks, validating the effectiveness of internalized comparative priors for robust, rank-invariant, and scale-steerable evaluation. The code is available at: https://github.com/SpatialVision-Research/SSAD_ACL2026_Findings.

1 Introduction

Subjective assessment is a fundamental capability of multimodal models: we routinely ask a model to judge whether an image is *beautiful*, *high-quality*, *safe*, or *appropriate*, and to output a score aligned with human preferences (Chen et al., 2024; Lee et al., 2024). Yet subjective scoring is inherently context-sensitive: the meaning of an absolute number is often underdetermined without a reference scale—humans rarely score in isolation, but place a target by comparing it against exemplars on an implicit ruler, *seeing what is good to know what is bad*. This suggests that the essence of subjective assessment is not pointwise regression, but comparative placement on a contextual ruler (Bradley and Terry, 1952; Plackett, 1975; Luce, 2012; Burges et al., 2005; Parikh and Grauman, 2011). This motivates a central question: **can we train multimodal judges to perform subjective assessment as comparative placement on a contextual ruler that is both rank-invariant and scale-steerable, rather than as isolated regression?**

A growing line of work improves subjective evaluation with reinforcement learning and preference modeling (Christiano et al., 2023; Ouyang et al., 2022; Rafailov et al., 2024). In image quality assessment (IQA), VisualQuality-R1 optimizes a vision-language judge with GRPO and Thurstone-style comparisons (Wu et al., 2025; Thurstone, 1927). Q-Insight extends this direction by combining scalar prediction with auxiliary degradation-understanding objectives (Li et al., 2025). In aesthetics assessment, Aes-R1 proposes a tailored RL algorithm and integrates absolute scoring with relative ranking signals (Liu et al., 2025). Despite strong empirical progress, many approaches still inject comparison indirectly: scalar scores are converted into pairwise/listwise rewards, and models can succeed by behaving like high-capacity regressors that fit dataset-specific mappings, rather than learning an explicit "ruler placement" strategy.

We propose **S^{2}AD**, a two-stage framework that operationalizes subjective assessment as **comparative analysis**, organized around the progression **from Mimesis to Metamorphosis**. The overall framework is shown in Figure 1. In **Stage 1 (Mimesis)**, the model learns by imitation, mirroring how humans are trained with exemplars. Concretely, for each query image, we retrieve a small reference set to establish a contextual ruler. We intro-

* Equal contribution.
† Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12957–12971
July 2-7, 2026

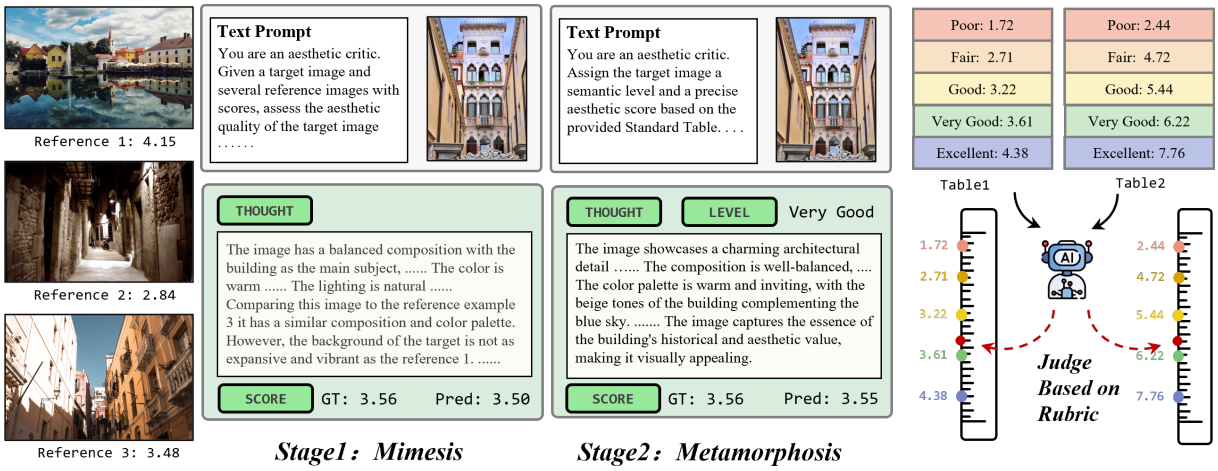

Figure 1: S^{2}AD overview. Stage 1 (Mimesis) trains a retrieval-augmented teacher to comparatively place a query image on a contextual ruler defined by scored reference images; Stage 2 (Metamorphosis) distills this comparative logic into retrieval-free Language Buttons with a Standard Table of level-to-score anchors, enabling rubric-aware and scale-steerable scoring without reference images at inference time.

duce **Dynamic Soft Positioning (DSP)** to train the teacher to place the query into the correct interval induced by these references. To prevent the judge from overfitting to a fixed numeric convention, we employ **Dynamic Rescale** to transform ground-truth scores into arbitrary, instance-specific scales. This forces the model to learn the logic of relative placement rather than memorizing absolute values, ensuring its comparative behavior remains robust across varying evaluation standards.

In **Stage 2 (Metamorphosis)**, the judge transitions from relying on external exemplars to carrying an internalized ruler. After learning an ordinal sense of quality through mimesis, the model employs **Language Buttons**—discrete semantic levels (e.g., `Poor` to `Excellent`)—to represent its internalized subjective priors. Reusing **Dynamic Rescale**, the model further decouples qualitative intuition from any fixed numeric range, and instantiates its judgments via a *Standard Table* (level-to-score anchors) that serves as an internal reference. Importantly, to align the student's *score logic* with the teacher's, we distill the teacher's comparison behavior onto a shared *virtual ruler* via a **Fidelity Reward**. This ensures the student preserves the learned placement strategy without external references at test time, enabling efficient and scale-agnostic inference.

Beyond accuracy and efficiency, S^{2}AD yields a practical byproduct: **scale-steerability**. Because Language Buttons decouple qualitative ranking from numeric conventions, the model can transfer across benchmarks by *anchoring* the same button levels to the target dataset's scoring convention. This is achieved via a minimal non-test *numeric anchoring* step, instantiating the button levels under the dataset's numeric scale without updating parameters. To establish the overall effectiveness of the framework, we evaluate S^{2}AD on aesthetic assessment as a representative subjective domain and further demonstrate transfer to other subjective evaluation settings, indicating that S^{2}AD is not tied to a single task.

In summary, our contributions are threefold.

- **Scale-aware explicit comparative training.** We introduce a reference-conditioned training paradigm and optimize DSP with Dynamic Rescale to supervise comparative placement on a contextual ruler, rather than pointwise regression.

- **Retrieval-free internalization via language buttons.** We compress reference-based judging into semantic priors by introducing Language Buttons and distilling the teacher's comparison behavior, enabling efficient inference without reference images.

- **Context steerability.** By grounding scoring in a controllable ruler, our framework supports coherent, rubric-aligned shifts of the output distribution under different scoring styles.

2 Related Work

Supervised MLLM judges for subjective visual assessment. Vision–language models (VLMs) and multimodal LLMs are increasingly explored as scalable judges for subjective evaluation, supporting rubric-based scoring and critique-style assessment in multimodal settings (Lee et al., 2024; Chen et al., 2024). A common early paradigm trains such judges primarily via supervised fine-tuning (SFT), prompting the model to produce scores under a fixed rubric (Lee et al., 2024; Cao et al., 2025). In the aesthetics domain, Cao et al. (2025) exemplifies this supervised line by fine-tuning an MLLM-based evaluator on expert-curated supervision for fine-grained aesthetic assessment. In quality assessment domain, recent MLLM-based IQA works further extend this supervised line in complementary directions: DepictQA moves beyond pure scalar prediction by enabling descriptive and comparative, language-based image quality assessment, while DeQA-Score focuses on accurate continuous score regression through score-distribution modeling and a Thurstone-style fidelity loss (You et al., 2024, 2025). While these approaches improve within-scale score accuracy or qualitative richness, neither line of work explicitly investigates rubric-shift robustness under heterogeneous numeric conventions, which is a central focus of our method.

Preference learning and reinforcement learning for subjective judges. Beyond supervised judge training, a parallel line of work leverages relative supervision through preference learning and reinforcement learning (RL), often improving ranking consistency and alignment with human taste. Reward modeling is a representative approach: ImageReward learns a text-to-image human preference reward model from expert comparisons and further uses it to optimize diffusion models, while HPS v2 provides a learned preference score trained on large-scale human pairwise choices as an automatic evaluation signal (Xu et al., 2023; Wu et al., 2023). Recently, efficient RL recipes such as Group Relative Policy Optimization (GRPO) have been developed and popularized in LLM post-training (Shao et al., 2024; Guo et al., 2025), and have begun to be adopted to train vision-language evaluators for image quality assessment (IQA) (Wu et al., 2025; Li et al., 2025). However, relative signals in prior RL setups are often used as a means to refine scalar predictors, instead of learning a context-aware placement rule from comparisons. As a result, models may generalize poorly when the numeric convention changes. In contrast, our Stage 1 explicitly casts subjective assessment as comparative analysis under a reference set.

Aesthetics/IQA datasets. Progress in subjective visual evaluation is driven by diverse datasets with heterogeneous annotation protocols and scoring ranges. For aesthetics, widely used benchmarks include AVA (1–10 ratings) (Murray et al., 2012), AADB (ratings with attribute annotations) (Kong et al., 2016), and datasets with dense annotations such as TAD66K (He et al., 2022); personalized benchmarks with rich attributes such as PARA further highlight preference-dependent scoring (Yang et al., 2022). Complementary aspect-focused benchmarks such as CADB emphasize composition assessment beyond holistic scores (Zhang et al., 2021). For image quality assessment (IQA), common testbeds include realistic smartphone photography datasets such as SPAQ and controlled distortion datasets such as KADID-10k (Fang et al., 2020; Lin et al., 2019). These dataset properties motivate studying scale-aware comparative supervision and controllable judging behavior across heterogeneous score conventions.

3 Methodology

In this section, we present S^{2}AD, a two-stage framework that operationalizes subjective assessment through a progression from *mimesis* to *metamorphosis*. As illustrated in Figure 2, the model evolves from explicit, reference-conditioned comparison to retrieval-free internalized judgment.

3.1 Problem Formulation

Given an input image $\mathcal{I}$ and an optional textual prompt $\mathcal{P}$, our goal is to predict a scalar aesthetic score $s \in [\mathcal{S}_{min}, \mathcal{S}_{max}]$ (e.g., $[1, 5]$). Since absolute aesthetic annotations can be subjective and exhibit non-trivial noise, we complement pointwise supervision with relative signals derived from comparisons, which often provide a more informative training signal for calibrating intervals and ordering. Motivated by this, we encourage the scorer to be consistent in its relative placement when compared against a small, context-aware anchor set.

Let $f_\theta(\mathcal{I}, \mathcal{P})$ denote a VLM parameterized by θ that outputs a score $\hat{s}$. For each query $(\mathcal{I}_q, \mathcal{P})$ with ground-truth score s_q^*, we dynamically construct an anchor set

$$\mathcal{M} = \left\{(\mathcal{I}_j, s_j^*)\right\}_{j=1}^{N}$$

Figure 2: Architecture of the proposed S^{2}AD framework. In Stage 1, the teacher model learns to position a query image relative to retrieved references. In Stage 2, the student model internalizes this capability via Language Buttons and distillation, enabling retrieval-free inference.

where $N = 4$. We optimize θ such that the predicted score $\hat{s}_q = f_\theta(\mathcal{I}_q, \mathcal{P})$ is consistent with s_q^* in terms of its relative position with respect to the anchor scores $\{s_j^*\}$.

Dynamic Rescale. To model heterogeneous scale conventions and prevent the scorer from overfitting to a fixed numeric ruler, we apply an instance-wise affine transformation to the score space:

$$\tilde{s} = a\, s + b \tag{1}$$

where a (scale) and b (shift) are sampled per instance. We apply the same (a, b) to the query and anchor scores within $\mathcal{M}$, yielding transformed scores $\tilde{s}_q^*$ and $\{\tilde{s}_j^*\}$ that define the instance-specific ruler. All comparison-based objectives in both Stage 1 and Stage 2 are computed in this transformed space, ensuring that learned comparative behavior remains informative under scale shifts.

3.2 Stage 1 (Mimesis): Retrieval-Augmented Explicit Comparative Training

As shown in Figure 2 (left), stage 1 (Mimesis) establishes a reference-conditioned training environment by leveraging an expert photo database categorized into aesthetic groups. By retrieving in-context reference samples that span different quality levels, we construct an explicit contextual ruler to supervise the teacher VLM. This allows the model to learn relative placement logic via Dynamic Soft Positioning (DSP) rather than fitting absolute numeric mappings.

Hybrid Multimodal Retrieval. We adopt hybrid retrieval (a weighted fusion of visual and textual similarity) and a score-bin coverage stratified sampling scheme to select N=4 reference images spanning different quality ranges as comparative anchors (details in Appendix A.1.1).

Comparison in the rescaled space. Following Eq. 1, we compute Stage 1 objectives on the transformed scores: the query has ground-truth $\tilde{s}_q^*$ and anchors have $\{\tilde{s}_j^*\}_{j=1}^N$. We denote the teacher prediction on this instance-specific ruler as $\hat{\tilde{s}}_q$.

Dynamic Soft Positioning. To supervise where the query should land on the ruler, we adopt Dynamic Soft Positioning (DSP), which maps a score to a soft position relative to $\mathcal{M}$:

$$\rho(\tilde{s}, \mathcal{M}) = 1 + \sum_{j=1}^{N} \sigma\big(\beta'(\tilde{s}_j^* - \tilde{s})\big) \tag{2}$$

$$\beta' = \frac{\beta}{a + \epsilon}, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, β' controls the sharpness and β is a hyperparameter. We then measure the placement error between the prediction and the ground truth:

$$\mathcal{L}_{\text{DSP}} = \Big|\rho(\hat{\tilde{s}}_q, \mathcal{M}) - \rho(\tilde{s}_q^*, \mathcal{M})\Big| \tag{4}$$

Finally, we convert it to a bounded reward:

$$\mathcal{R}_{\text{rank}} = \exp(-\lambda \mathcal{L}_{\text{DSP}}) \tag{5}$$

where λ is a scaling hyperparameter.

Interval consistency. While $\mathcal{R}_{\text{rank}}$ supervises the global relative position, we additionally encourage local interval calibration between the two nearest reference scores that bracket the query. Let the reference scores in $\mathcal{M}$ be sorted as $\tilde{s}_{(1)}^* \le \cdots \le \tilde{s}_{(N)}^*$.

Denote by $\tilde{s}_l^*$ and $\tilde{s}_r^*$ the immediate left/right neighbors that bracket the ground-truth $\tilde{s}_q^*$ in this ordered list. We define a normalized interval coordinate

$$\kappa(\tilde{s}, \mathcal{M}) = \frac{\tilde{s} - \tilde{s}_l^*}{\tilde{s}_r^* - \tilde{s}_l^* + \epsilon} \tag{6}$$

Then the interval discrepancy is

$$\mathcal{R}_{\text{int}} = \exp\left(-\gamma \left| \kappa(\hat{\tilde{s}}_q, \mathcal{M}) - \kappa(\tilde{s}_q^*, \mathcal{M}) \right| \right) \tag{7}$$

where γ is a scaling hyperparameter. Finally, we combine the two signals into the Stage 1 reward:

$$\mathcal{R}_{\text{Stage 1}} = \tfrac{1}{2}\mathcal{R}_{\text{rank}} + \tfrac{1}{2}\mathcal{R}_{\text{rank}}\mathcal{R}_{\text{int}} \tag{8}$$

We combine $\mathcal{R}_{\text{rank}}$ and $\mathcal{R}_{\text{int}}$ in a coarse-to-fine manner. Intuitively, $\mathcal{R}_{\text{rank}}$ captures the global ordering of the query relative to the ruler, while $\mathcal{R}_{\text{int}}$ refines the *local* calibration within the bracketing interval. The multiplicative term $\mathcal{R}_{\text{rank}}\mathcal{R}_{\text{int}}$ acts as a soft gate: interval-level rewards are emphasized only when the model has already placed the query at an approximately correct global position.

Optimization. We optimize the Stage 1 teacher with GRPO using the reward $\mathcal{R}_{\text{Stage 1}}$ computed on the retrieval-augmented ruler.

3.3 Stage 2: Metamorphosis — Retrieval-Free Internalization with Language Levels

As shown in Figure 2 (right), stage 2 freezes the Stage-1 trained teacher and optimizes a retrieval-free student (initialized from the teacher model) to score a query image alone at test time; the student internalizes the comparative logic via a fixed Language-Button prompt and distillation rewards that align its placement behavior with the teacher.

Standard prompt with five semantic levels. We introduce a lightweight standard prompt that defines five discrete Language Buttons (`Poor / Fair / Good / Very Good / Excellent`). The standard table is constructed by partitioning the training-score distribution into five percentile-aligned segments, and assigning each level an approximate target computed from the corresponding segment. Details, templates, and numeric values are provided in Appendix A.2.1.

Fidelity Reward via Bhattacharyya Coefficient. To encourage distributional consistency between the student prediction and the teacher signal, we introduce a Fidelity Reward that aligns their pairwise comparison profiles with respect to the same reference set. Given a teacher score $\hat{\tilde{s}}_t$ and a student-predicted score $\hat{\tilde{s}}_s$, we compute the probability that each score exceeds a reference score $\tilde{s}^*$ using a temperature-controlled sigmoid:

$$P_{\mathcal{T}}^k = \sigma\left(\frac{\hat{\tilde{s}}_t - \tilde{s}^*}{\tau}\right), \quad P_{\mathcal{S}}^k = \sigma\left(\frac{\hat{\tilde{s}}_s - \tilde{s}^*}{\tau}\right) \tag{9}$$

where τ controls comparison sharpness. We then measure the affinity between the teacher and student Bernoulli distributions for each reference via the Bhattacharyya Coefficient, and average over all references:

$$\mathcal{R}_{\text{f}} = \frac{1}{K}\sum_{k=1}^{K}\left(\sqrt{P_{\mathcal{T}}^k P_{\mathcal{S}}^k} + \sqrt{(1-P_{\mathcal{T}}^k)(1-P_{\mathcal{S}}^k)}\right) \tag{10}$$

This reward encourages the student to match the teacher not only in absolute score magnitude, but also in how the score is positioned relative to multiple references, yielding a calibrated and scale-aware distillation signal.

Virtual DSP for retrieval-free placement. To complement Fidelity, we reuse the DSP computation from Stage 1 to provide direct rank-aware supervision. We substitute the visual ruler $\mathcal{M}$ with uniform anchors U, obtained by sampling scalar points across the valid score range (e.g., $[1, 5]$ for PARA). The student's relative positioning is supervised by matching its placement on U to the ground-truth placement:

$$\mathcal{L}_{\text{vDSP}} = \left|\rho(\tilde{s}_s, U) - \rho(\tilde{s}^*, U)\right| \tag{11}$$

This discrepancy is converted into $\mathcal{R}$rank using Eq. 5. In practice, these uniform anchors are mapped into the current rescaled range to maintain consistency with our Dynamic Rescale strategy. Details of how to select uniform anchors are provided in Appendix A.2.3.

We optimize the student with GRPO under a multi-objective reward:

$$\mathcal{R}_{\text{total}} = w_1\mathcal{R}_f + w_2\mathcal{R}_{\text{rank}} + w_3\mathcal{R}_{\text{level}}. \tag{12}$$

Here $\mathcal{R}_{\text{level}}$ is a lightweight semantic consistency reward that encourages the predicted level to match the ground-truth level ℓ^* induced by the percentile-aligned table:

$$\mathcal{R}_{\text{level}} = \mathbf{I}[\hat{\ell} = \ell^*]. \tag{13}$$

Auxiliary stabilizers such as output-format constraints and range guards are used in practice.

4 Experiments Setup

4.1 Implementation Details

Benchmarks and Metrics. We employ PARA (Yang et al., 2022) as the in-domain training data set of our method. To evaluate the effectiveness and generalization, we conduct extensive testing on six additional benchmarks spanning aesthetic assessment and image quality assessment (IQA). For out-of-distribution (OOD) aesthetic evaluation, we test on AVA (Murray et al., 2012), TAD66K (He et al., 2022), CADB (Zhang et al., 2021), and AADB (Kong et al., 2016). To further examine cross-task generalization beyond aesthetics, we report results on two widely used IQA datasets: KADID-10k (Lin et al., 2019) and SPAQ (Fang et al., 2020). We report Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) between predicted scores and ground-truth scores: SRCC measures ranking consistency and is the primary metric under rubric/scale shifts, while PLCC reflects linear alignment under a given numeric convention.

Baselines. We compare S^{2}AD against four categories of models: (i) open-source VLMs, including Qwen2.5-VL (7B/72B) (Bai et al., 2025), InternVL3 (8B/38B) (Zhu et al., 2025), and LLaVA-OneVision (Li et al., 2024); (ii) Task-specialized scorers, such as ArtiMuse (Cao et al., 2025) for aesthetics, and Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025) for IQA; (iii) Closed-source reference, represented by GPT-4o (OpenAI et al., 2024) and (iv) Traditional supervised model, Charm (Behrad et al., 2025), a classical supervised aesthetic predictor trained on PARA, included as a non-VLM reference point.

Training Details. We use Qwen2.5-VL-7B (Bai et al., 2025) as the backbone. Stage 1 fine-tunes the pretrained backbone to obtain a reference-conditioned judge. Stage 2 then continues training from the Stage 1 checkpoint to internalize the comparative ruler into language buttons. In both stages, we fine-tune all model parameters. All experiments are conducted on 8×A100 GPUs (80GB each) with a global batch size of 32 and a learning rate of 2e−6, and we train for one epoch per stage. For GRPO, the group size is G=8 in Stage 1 and G=12 in Stage 2. (details in Appendix B).

4.2 Evaluation protocol

We explicitly specify the inference-time context used by each stage.

Stage 1. Although Stage 1 can be paired with retrieval, we adopt a deployment-oriented evaluation that avoids per-query retrieval at test time. For each target dataset, we *sample once* a fixed N=4 images from the training split, such that their ground-truth scores span low-to-high. Every test image is then evaluated under the same four-image context.

Stage 2. Stage 2 is retrieval-free by design, but transferring to a new dataset requires aligning numeric outputs to the target dataset's score convention. We therefore provide a minimal amount of *scale information* using a small non-test calibration split. Concretely, for each target dataset, we sample a calibration set $\mathcal{C}$ of size M=200 from the **non-test split** (training/validation). We ask the model to assign each image $x \in \mathcal{C}$ to one of five semantic levels (`Poor / Fair / Good / Very Good / Excellent`), and compute each level's numeric anchor as the mean ground-truth score of images assigned to that level:

$$a_l = \mathbb{E}_{x \in \mathcal{C},\, \hat{l}(x)=l}[s_{\mathrm{GT}}(x)]. \tag{14}$$

The resulting anchor table $\{a_l\}_{l=1}^{5}$ is fixed and used for scoring all test images. This step uses no test images or test labels and does not update model parameters; it only calibrates the numeric scale required for cross-dataset evaluation.

5 Results and Analysis

5.1 Main Results on Aesthetic Assessment

As shown in Table 1, S^{2}AD consistently achieves state-of-the-art performance across all OOD benchmarks. Stage 1 demonstrates superior OOD robustness by leveraging explicit exemplars, while Stage 2 maintains high competitive correlations on PARA (in-domain) without any retrieval overhead. Notably, both stages significantly outperform proprietary models like GPT-4o, validating the effectiveness of our comparative training. Although Charm, as a classical non-VLM supervised baseline trained directly on PARA human ratings, achieves higher in-domain correlation on PARA, it falls behind S^{2}AD on every OOD benchmark. This suggests that our comparative training strategy transfers better across datasets than conventional dataset-specific supervised fitting.

Method	In-domain	Out-of-distribution				Average
	PARA	AVA	CADB	TAD66K	AADB	
SRCC						
Qwen2.5VL-7B (Bai et al., 2025)	0.6702	0.3518	0.5442	0.2242	0.5474	0.4676
Qwen2.5VL-72B (Bai et al., 2025)	0.6978	0.4179	0.5859	0.2397	0.5614	0.5005
InternVL3-8B (Zhu et al., 2025)	0.6874	0.3667	0.5272	0.2037	0.5949	0.4760
InternVL3-38B (Zhu et al., 2025)	0.7070	0.3950	0.5343	0.2263	0.5713	0.4868
LLaVA-OneV-8B (Li et al., 2024)	0.7225	0.4188	0.5103	0.2391	0.4137	0.4609
Q-Insight (Li et al., 2025)	0.7683	0.4022	0.4954	0.2008	0.5832	0.4900
Artimuse (Cao et al., 2025)	0.5565	0.4260	0.2955	0.2401	0.3600	0.3756
Visualquality-r1 (Wu et al., 2025)	0.6863	0.4148	0.4724	0.1876	0.5086	0.4539
GPT-4o (OpenAI et al., 2024)	0.6784	*0.5013*	0.5380	0.2515	0.5317	0.5002
Charm† (Behrad et al., 2025)	**0.9050**	0.4210	0.6024	0.2802	0.6297	0.5677
Ours (Stage 1)	0.8396	**0.5226**	*0.6037*	**0.3229**	**0.6828**	**0.5943**
Ours (Stage 2)	*0.8805*	0.4884	**0.6122**	*0.2899*	*0.6811*	*0.5904*
PLCC						
Qwen2.5VL-7B (Bai et al., 2025)	0.7528	0.3684	0.5319	0.2282	0.4933	0.4749
Qwen2.5VL-72B (Bai et al., 2025)	0.7329	0.3992	0.5271	0.2413	0.5134	0.4828
InternVL3-8B (Zhu et al., 2025)	0.7767	0.2669	0.5242	0.1941	0.5587	0.4641
InternVL3-38B (Zhu et al., 2025)	0.7616	0.3551	0.5160	0.2317	0.5123	0.4753
LLaVA-OneV-8B (Li et al., 2024)	0.7578	0.4069	0.4388	0.2337	0.3961	0.4467
Q-Insight (Li et al., 2025)	0.8273	0.3925	0.4539	0.2173	0.4909	0.4764
Artimuse (Cao et al., 2025)	0.6681	0.4252	0.3087	0.2357	0.3423	0.3960
Visualquality-r1 (Wu et al., 2025)	0.7877	0.4197	0.4598	0.2068	0.4832	0.4714
GPT-4o (OpenAI et al., 2024)	0.7382	0.4283	0.5172	0.2394	0.5239	0.4894
Charm† (Behrad et al., 2025)	**0.9380**	0.4141	0.6056	0.2849	0.6301	0.5746
Ours (Stage 1)	0.8941	**0.5232**	*0.6059*	**0.3501**	*0.6680*	**0.6082**
Ours (Stage 2)	*0.9146*	*0.5071*	**0.6129**	*0.3103*	**0.6824**	*0.6055*

Table 1: Main results on in-domain and out-of-distribution aesthetic datasets. Our models are trained on PARA only, The top two results are highlighted in bold and italic underline. Charm† is traditional supervised aesthetic predictor trained on PARA.

Fairness check. To rule out that our gains stem from extra scale information or reference access, we apply the same protocols to some baseline judges. (1) **RAG fairness:** we provide the same retrieved reference context to baseline judges (*With RAG*) and compare against their original inference (*No-RAG*). (2) **Rubric calibration fairness:** we apply the same non-test numeric anchoring to baseline judges (*With rubric calibration*) and compare against their original predictions (*No-Table*). As shown in Table 2 (Details in Appendix C.1), these protocols do not yield consistent improvements for baselines and often degrade correlation, whereas S^2AD benefits from both, indicating that non-test anchoring and reference access are not free performance boosters; they become effective when the judge has learned a transferable comparative ruler.

5.2 Rubric/Table Shift Robustness

A key challenge in subjective assessment is *rubric/table shift*: the same semantic quality can correspond to different numeric targets under different scoring conventions. A reliable judge should preserve **rubric-invariant ranking** while allowing **rubric-dependent calibration** according to the specified table.

Protocol. We conduct a controlled stress test on a subset of 500 randomly selected images from the AVA dataset . We consider several anchor tables that differ in range and location (Default, Wide-Range, Compressed, High-Shift), and report SRCC under each setting.

Rank stability under rubric shifts. As shown in Figure 3, S^2AD maintains nearly unchanged SRCC across different anchor tables, indicating that the

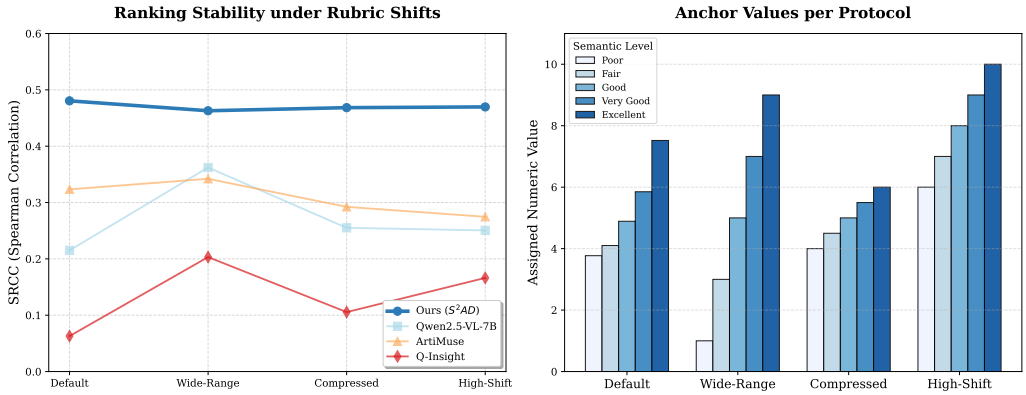

Figure 3: **Rubric/table shift robustness.** Left: SRCC across different rubric tables. Right: the corresponding anchor tables used in prompts, illustrating how numeric conventions differ in range and location.

Method	PARA		CADB	
	SRCC	PLCC	SRCC	PLCC
ΔRAG (With RAG − No-RAG)				
Qwen2.5VL-7B	-0.0062	+0.0209	-0.1531	-0.1214
Q-Insight	-0.0369	-0.0007	+0.0138	+0.0269
ArtiMuse	-0.0231	-0.0946	+0.0136	+0.0010
Ours	**+0.0327**	**+0.0338**	**+0.0483**	**+0.0552**
ΔCalib (With rubric calibration − No-Table)				
Qwen2.5VL-7B	-0.1562	-0.1054	-0.2190	-0.2691
Q-Insight	-0.0184	-0.0303	-0.0446	-0.0062
ArtiMuse	-0.0535	-0.1321	-0.0324	-0.0925
Ours	**+0.0190**	**+0.0126**	**+0.0171**	**+0.0310**

Table 2: **Fairness check.** We report performance deltas for (i) adding retrieved references (**ΔRAG**) and (ii) applying non-test rubric calibration(**ΔCalib**).

learned preference ordering is largely independent of the specific numeric convention. In contrast, several baselines exhibit substantially larger SRCC volatility under the same prompt-only table shift, suggesting that their ranking behavior is more entangled with a particular rubric. Additional statistics on mean/std shifts and distributional analyses are provided in Appendix C.2.

5.3 Effect of Context, Anchoring, and Training Ablations

In this section, we analyze how inference-time context, evaluation-time anchoring, and training variants affect performance.

Necessity of the contextual ruler. As shown in Table 3, removing the retrieved reference pack during Stage 1 inference (Stage 1 w/o RAG) consistently degrades performance (avg. SRCC: 0.5943 → 0.5268), indicating that the teacher benefits from explicit anchors for comparative placement.

Efficacy of internalization. Table 3 also shows that disabling the non-test numeric anchoring in Stage 2 (Stage 2 w/o Table) causes only a mild drop, while maintaining strong correlation (avg. SRCC: 0.5796). This suggests that Language Buttons capture most of the internalized comparative logic, while anchoring mainly refines the dataset-specific numeric mapping.

Two-stage synergy. As further shown in Table 3, training Stage 2 directly from the backbone without the Stage 1 teacher and Fidelity (Stage 2-only w/o Fidelity) leads to a large degradation (avg. SRCC: 0.5218), supporting that metamorphosis depends on both Stage 1 imitation priors and Fidelity-based behavioral distillation.

Sensitivity to reference count (N). As described in Section 3.2 and Appendix A.1.1, Stage 1 constructs a contextual ruler via score-bin coverage, sampling one reference per bin with $N{=}B$ by construction. We further ablate $N \in \{2, 4, 6, 8\}$ on four OOD benchmarks, with results shown in Table 4. $N{=}2$ consistently underperforms, suggesting that too few references provide insufficient ruler coverage. $N{=}4$ and $N{=}6$ achieve competitive results across datasets, while increasing to $N{=}8$ provides no further gain. We therefore adopt $N{=}B{=}4$, which offers strong performance while keeping VRAM overhead manageable.

5.4 Generalization to IQA

To test generalization beyond aesthetics, we evaluate zero-shot transfer to two IQA benchmarks, KADID-10k and SPAQ (Table 5). Despite the domain shift (aesthetics → distortion-aware quality), S^2AD consistently improves over the general VLM baseline, suggesting the learned scale-aware com-

Method	In-domain	Out-of-distribution				Average
	PARA	AVA	CADB	TAD66K	AADB	
SRCC						
Stage 1(w/o RAG)	0.8069	0.4501	0.5554	0.2453	0.5762	0.5268
Stage 1	**0.8396**	**0.5226**	**0.6037**	**0.3229**	**0.6828**	**0.5943**
Stage 2(w/o Table)	0.8615	0.4764	0.5951	0.2820	0.6828	0.5796
Stage 2-only(w/o Fidelity)	0.8043	0.4351	0.5602	0.2410	0.5684	0.5218
Stage 2	**0.8805**	**0.4884**	**0.6122**	**0.2899**	**0.6811**	**0.5904**
PLCC						
Stage 1(w/o RAG)	0.8603	0.4604	0.5507	0.2612	0.5450	0.5355
Stage 1	**0.8941**	**0.5232**	**0.6059**	**0.3501**	**0.6680**	**0.6082**
Stage 2(w/o Table)	0.9020	0.4806	0.5819	0.2861	0.6681	0.5837
Stage 2-only(w/o Fidelity)	0.8387	0.4340	0.5536	0.2552	0.4930	0.5149
Stage 2	**0.9146**	**0.5071**	**0.6129**	**0.3103**	**0.6824**	**0.6055**

Table 3: Ablation studies on context, anchoring, and training pipeline. The best results are highlighted in bold.

Dataset	N=2	N=4	N=6	N=8
SRCC				
AADB	0.658	**0.683**	0.663	0.678
AVA	0.504	**0.523**	0.499	0.497
CADB	0.591	0.604	**0.612**	0.610
TAD66K	0.300	0.323	**0.336**	0.317
PLCC				
AADB	0.632	**0.668**	0.652	0.661
AVA	0.513	**0.523**	0.496	0.502
CADB	0.579	0.606	**0.620**	0.607
TAD66K	0.322	0.350	**0.353**	0.335

Table 4: Stage 1 SRCC and PLCC under varying reference counts N. Best results per dataset are in bold.

Method	KADID		SPAQ	
	SRCC	PLCC	SRCC	PLCC
Q-Insight†	**0.7331**	0.6709	**0.9034**	**0.9056**
Visualquality-r1†	0.7110	0.6760	0.9011	0.9017
Qwen2.5VL-7B	0.5601	0.5575	0.8595	0.8564
Ours (Stage 1)	0.6731	0.6620	0.8800	0.8849
Ours (Stage 2)	0.6482	**0.6784**	0.8859	0.8816

Table 5: Results on IQA datasets. The best results are highlighted in bold. † indicates that the model was trained on the KADID and SPAQ datasets.

parative ruler transfers to broader subjective assessment. Note that methods marked with † are trained on the target IQA dataset.

6 Conclusion

We presented **S^{2}AD**, a two-stage framework that reframes subjective assessment as comparative placement on a contextual ruler, rather than pointwise regression. The framework evolves from explicit visual comparisons to internalized semantic anchors, enabling efficient, rank-invariant, and scale-steerable evaluation without needing reference images at inference. Across multiple aesthetics benchmarks and transfer experiments on IQA datasets, S^{2}AD achieves strong correlation with human labels and exhibits robustness under rubric/table shifts. In future work, we plan to extend S^{2}AD to broader subjective settings (e.g., safety, appropriateness, multi-attribute rubrics), explore richer internal anchor structures beyond discrete levels.

Limitations

Scope of Subjective Attributes. While our framework demonstrates state-of-the-art results in aesthetic assessment and competitive performance on image quality assessment (IQA), subjective evaluation is an expansive field that encompasses diverse dimensions such as image safety, humor, and emotional resonance. Our current study primarily focuses on holistic quality and aesthetic priors. Extending the "Metamorphosis" framework to these highly specialized or multi-dimensional subjective domains remains an open area for investigation.

Granularity of internalized anchors. We operationalize Stage 2 (Metamorphosis) using five discrete semantic levels as "Language Buttons" to represent the internalized ruler. Although this configuration yields strong performance and aligns with standard human grading scales, certain extremely fine-grained evaluation tasks might benefit from a higher density of semantic anchors or a continuous latent representation. We leave the optimization of anchor density for specific niche applications to future studies.

Training-time retrieval bias. Stage 1 relies on retrieved reference images to construct the con-

textual ruler, so its behavior can be affected by biases in the retrieval pipeline, such as similarity mismatches. In this work, we partially mitigate this risk through score-bin coverage sampling, hybrid visual-text retrieval, and randomized top-K selection, which improve ruler diversity. More systematic mitigation remains future work, such as stronger diversity-aware retrieval.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. 2025. Charm: The missing piece in vit fine-tuning for image aesthetic assessment. *Preprint*, arXiv:2504.02522.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4).

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

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. 2025. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *Preprint*, arXiv:2507.14533.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. 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.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. 2022. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 942–948. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision – ECCV 2016*, pages 662–679, Cham. Springer International Publishing.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *Preprint*, arXiv:2401.06591.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

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

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3.

Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. 2025. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. *Preprint*, arXiv:2509.21871.

R. Duncan. Luce. 2012. Individual choice behavior : A theoretical analysis.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415.

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, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510.

R. L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

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

L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. 2025. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Preprint*, arXiv:2505.14460.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Preprint*, arXiv:2304.05977.

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

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

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. 2024. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *Preprint*, arXiv:2312.08962.

Bo Zhang, Li Niu, and Liqing Zhang. 2021. Image composition assessment with saliency-augmented multi-pattern pooling. In *British Machine Vision Conference*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Details of S2AD

A.1 Stage 1 (Mimesis)

A.1.1 Hybrid Retrieval and Score-Bin Coverage Sampling

This appendix details how we construct the reference pack $\mathcal{M}$ for Stage 1 via hybrid retrieval and score-bin coverage, as referenced in Section 3.2.

Hybrid multimodal similarity. For each image, we extract a CLIP image embedding $v \in \mathbb{R}^d$ and a CLIP text embedding $t \in \mathbb{R}^d$ computed from a short VLM-generated caption. All embeddings are ℓ_2-normalized, so inner product equals cosine similarity. Given a query image $\mathcal{I}_q$ with embeddings (v_q, t_q) and a candidate image $\mathcal{I}_c$ with (v_c, t_c), we compute a hybrid similarity:

$$\mathrm{Sim}(\mathcal{I}_q, \mathcal{I}_c) = \lambda_v \langle v_q, v_c \rangle + \lambda_t \langle t_q, t_c \rangle \qquad (15)$$

where λ_v and λ_t control the relative importance of visual and textual cues.

Score-bin coverage sampling (N=4). To ensure the reference pack spans diverse quality ranges, we stratify candidates by their ground-truth aesthetic score into B bins (uniform bins over the training score range; in practice B=4 with $[1, 2), [2, 3), [3, 4), [4, 5)$). For each bin b, we restrict the candidate pool to images whose scores fall into that bin and rank them by $\mathrm{Sim}(\mathcal{I}_q, \mathcal{I}_c)$. To avoid repeatedly selecting near-duplicate anchors, we perform randomized top-K sampling: we take the top-K candidates within each bin and uniformly sample one as the anchor for that bin. The final reference pack $\mathcal{M}$ contains one anchor per bin, yielding N=4 references per query.

Table 6: Structured prompt used in Stage 1 (reference-conditioned).

Prompt Text

You are an expert aesthetic critic capable of adapting to arbitrary scoring scales based on context. Given a **target image** and several **reference images with scores**, assess the aesthetic quality of the target image *relative to the references* and output a scalar score on the **current scale**.

Current scale: range [Min, Max] (Worst → Best).

Reference examples:
(1) **Reference Image 1** with **Score 1**
(2) **Reference Image 2** with **Score 2**
(3) **Reference Image 3** with **Score 3**
(4) **Reference Image 4** with **Score 4**

Target: Target Image

First output the thinking process in `<think>...</think>` tags, explicitly analyzing *composition, color/tone, lighting, depth of field, and content/semantics*, and comparing the target image with the references. Then output the final score with only one number in `<answer>...</answer>` tags.

Required output format:
`<think> ... </think>`
`<answer> NUMBER </answer>`

A.1.2 Stage 1 Prompt Template

Table 6 shows the structured text prompt used in Stage 1. At training time, the query is augmented with a reference pack of $N{=}4$ images and their (rescaled) scores, which define the instance-specific ruler; the model is required to produce a rationale in `<think>` and a single scalar score in `<answer>`.

A.2 Stage 2 (Metamorphosis)

A.2.1 Standard Table Construction

As described in Section 3.3, we build a Standard Table for five semantic levels (`Poor / Fair / Good / Very Good / Excellent`) from the training-score distribution. We compute percentile cut points at $p \in \{20, 40, 60, 80, 100\}$ and form five contiguous intervals:

$$\begin{aligned} &\mathcal{B}_1 = [Q_0,\, Q_{20}), \quad \mathcal{B}_2 = [Q_{20},\, Q_{40}), \\ &\mathcal{B}_3 = [Q_{40},\, Q_{60}), \quad \mathcal{B}_4 = [Q_{60},\, Q_{80}), \\ &\mathcal{B}_5 = [Q_{80},\, Q_{100}] \end{aligned} \tag{16}$$

where Q_p denotes the p-th percentile of training scores (with Q_0 being the minimum). For numerical stability, we optionally allow a small upper slack on Q_{100} when defining the boundary set.

Anchor values. We assign each level an *anchor* (an "around" value) as the midpoint of its interval: for $\mathcal{B}_i = [\ell_i, u_i]$, we set $a_i = (\ell_i + u_i)/2$.

Example. On PARA (raw score range $[1, 5]$), one instantiation is

$$\mathbf{b} \approx [1.00,\ 2.44,\ 2.98,\ 3.45,\ 3.76,\ 5.00] \tag{17}$$

which yields midpoint anchors

$$\mathbf{a} \approx [1.72,\ 2.71,\ 3.22,\ 3.61,\ 4.38] \tag{18}$$

corresponding to (`Poor / Fair / Good / Very Good / Excellent`), respectively. Stage 2 operates in the rescaled space $\tilde{s} = as + b$, so we apply the same affine transform to $\mathbf{a}$ before inserting the table into the prompt.

A.2.2 Stage 2 Prompt Template

Table 7 shows the structured prompt used in Stage 2, where visual references are removed. Instead, the model is provided with an anchor-only Standard Table (five semantic levels with "around" values) and is asked to (i) output a discrete level in `<level>` and (ii) output a scalar score in `<answer>`, with intermediate rationale in `<think>`.

A.2.3 Virtual DSP rank reward $\mathcal{R}_{\text{rank}}$.

To compute the Stage 2 DSP-style rank reward without visual references, we substitute $\mathcal{M}$ with uniform anchors U sampled across the valid score range (mapped into the current rescaled convention). In implementation, we use 11 evenly spaced base anchors over $[1, 5]$:

$$\begin{aligned} U_{\text{base}} = \{&1.0, 1.4, 1.8, 2.2, 2.6, 3.0, \\ &3.4, 3.8, 4.2, 4.6, 5.0\} \end{aligned} \tag{19}$$

and map them via the same affine transform to obtain anchors in the current rescaled ruler.

A.3 Dynamic Rescale Sampling

We apply instance-wise affine transforms $\tilde{s} = as + b$. In our implementation, a and b are sampled from discrete sets to provide stable training signals while covering a wide range of scale and shift patterns:

$$\begin{aligned} &a \in \{\,0.1,\ 0.2,\ \ldots,\ 1.0\,\} \cup \{\,2,\ 3,\ \ldots,\ 10\,\}, \\ &b \in \{\,-10,\ -9,\ \ldots,\ 9,\ 10\,\} \end{aligned} \tag{20}$$

The same (a, b) is applied to both query and reference scores within the same instance.

Table 7: Structured prompt used in Stage 2 (anchor-only variant shown).

Prompt Text

You are an expert aesthetic critic. Your task is to assign the target image a semantic level and a precise aesthetic score based on the provided Standard Table.
(1) First, output the semantic level in <level> </level> tags.
(2) Then, provide a detailed rationale in <think> </think> tags.
(3) Finally, output the score in <answer> </answer> tags.

Input: Target Image

Evaluate based on the standard table below:

Standard Table:
- <Poor>: around Anchor_Poor
- <Fair>: around Anchor_Fair
- <Good>: around Anchor_Good
- <Very Good>: around Anchor_VeryGood
- <Excellent>: around Anchor_Excellent

Required output format:
<level> Poor/Fair/Good/Very Good/Excellent </level>
<think> ... </think>
<answer> NUMBER </answer>

B Implementation Details

We use Qwen2.5VL-7B as the backbone. Stage 1 fine-tunes the pretrained backbone to obtain a reference-conditioned judge. Stage 2 then continues training from the Stage 1 checkpoint. We perform full parameter fine-tuning in both stages.

B.1 Model and Infrastructure

- **Hardware:** All experiments are conducted on a cluster of 8×A100 (80GB) GPUs.
- **Optimization:** We use a global batch size of 32 and a learning rate of 2×10^{-6}, training for one epoch per stage.
- **Dataset:** We use a subset of the PARA training set for model training, which consists of 18,980 images.

B.2 Stage-specific Training Protocols

We optimize the vision-language judge using **Group Relative Policy Optimization (GRPO)**.

- **Group Size:** We set the group size to $G = 8$ for Stage 1 (Mimesis) and $G = 12$ for Stage 2 (Metamorphosis). The group size in Stage 1 is smaller than that in Stage 2 due to computational resource constraints: Stage 1 involves multi-image reasoning with a reference pack, which requires significantly more VRAM per generation than the retrieval-free inference used in Stage 2.
- **Stage 1 Retrieval:** For the hybrid multimodal retrieval (Appendix A.1), we employ visual and textual similarity weights $(\lambda_v, \lambda_t) = (0.4, 0.6)$. We apply randomized top-K sampling with $K = 10$ to select $N = 4$ reference images, ensuring each anchor represents one distinct score bin.
- **Stage 2 Optimization:** In the internalization stage (Appendix A.2), the multi-objective reward weights in Eq. (12) are set to $(w_1, w_2, w_3) = (1.0, 1.0, 0.5)$. For the virtual DSP rank reward, we define $m = 11$ uniform anchors spanning the rescaled score range.

C Expanded Experimental Results

C.1 Fairness Check – Full Results

This section provides the full results underlying the fairness check reported in Table 2 of the main text. We evaluate two protocols on representative datasets (PARA and CADB) and report both SRCC and PLCC. **(i) RAG fairness:** we provide baselines with the same fixed reference pack used by our Stage 1 evaluation (*With RAG*) and compare it against the baseline's original inference without references (*No-RAG*); the absolute results are shown in Table 8. **(ii) Rubric calibration fairness:** we apply the same non-test anchor-table calibration used by our Stage 2 transfer (*With Table*) and compare it against the baseline's original scoring prompt without an anchor table (*No-Table*); the absolute results are shown in Table 9.

C.2 Additional Statistics for Rubric/Table Shift Robustness

This section complements the rubric/table-shift robustness study in the main paper by reporting additional distributional statistics of the predicted

Method	PARA		CADB	
	SRCC	PLCC	SRCC	PLCC
Qwen2.5VL-7B-RAG	0.664	0.774	0.391	0.411
Qwen2.5VL-7B	0.670	0.753	0.544	0.532
Q-Insight-RAG	0.731	0.827	0.509	0.481
Q-Insight	0.768	0.827	0.495	0.454
Artimuse-RAG	0.533	0.573	0.309	0.310
Artimuse	0.556	0.668	0.295	0.309
Stage 1	**0.840**	**0.894**	**0.604**	**0.606**
Stage 1 (w/o RAG)	0.807	0.860	0.555	0.551

Table 8: Fairness check (RAG access) — full results on PARA and CADB. We report SRCC/PLCC when provided with the same fixed reference pack at inference time (With RAG) versus its original reference-free setting (No-RAG).

Method	PARA		CADB	
	SRCC	PLCC	SRCC	PLCC
Qwen2.5VL-7B-Table	0.514	0.647	0.325	0.263
Qwen2.5VL-7B	0.670	0.753	0.544	0.532
Q-Insight-Table	0.750	0.797	0.451	0.448
Q-Insight	0.768	0.827	0.495	0.454
Artimuse-Table	0.503	0.536	0.263	0.217
Artimuse	0.556	0.668	0.295	0.309
Stage 2	**0.881**	**0.915**	**0.612**	**0.613**
Stage 2 (w/o Table)	0.862	0.902	0.595	0.582

Table 9: Fairness check (non-test rubric calibration) — full results on PARA and CADB. We report SRCC/PLCC with our non-test anchor-table calibration (With Table) versus the original setting without an anchor table (No-Table); this calibration uses no test images/labels and updates no parameters.

scores under controlled changes of the anchor table. Our goal is to verify not only that a judge preserves *ranking* under rubric shifts, but also that it exhibits a *coherent and predictable* change in its numeric outputs when the scoring convention is modified.

Controlled anchor tables. We construct four anchor tables (Table 10) to simulate common rubric variations: (i) a default table, (ii) a wider-range table (stretching the numeric span), (iii) a compressed table (shrinking the numeric span), and (iv) a high-shift table (overall upward shift). All tables share the same five semantic levels (`Poor / Fair / Good / Very Good / Excellent`) but differ in their numeric anchors, thus isolating rubric/table shift from other confounders.

Distribution	Poor	Fair	Good	Very Good	Excellent
Default	3.77	4.10	4.89	5.85	7.52
Wide_Range	1.00	3.00	5.00	7.00	9.00
Compressed	4.00	4.50	5.00	5.50	6.00
High_Shift	6.00	7.00	8.00	9.00	10.00

Table 10: Anchor tables (rubrics) used in the controlled rubric/table-shift robustness test.

Rubric	SRCC	Mean	Std
Qwen2.5VL-7B			
Default	0.215	4.13	1.78
Wide-Range	0.362	5.01	2.32
Compressed	0.255	4.13	1.98
High_Shift	0.251	6.01	3.52
Q-Insight			
Default	0.063	2.82	2.23
Wide-Range	0.203	4.45	2.56
Compressed	0.106	3.44	2.26
High_Shift	0.166	5.49	3.67
Artimuse			
Default	0.323	4.34	1.02
Wide-Range	0.342	3.90	2.15
Compressed	0.292	4.51	0.73
High_Shift	0.275	6.15	1.72
Ours			
Default	0.481	5.04	0.65
Wide-Range	0.463	5.40	1.80
Compressed	0.468	5.05	0.44
High_Shift	0.470	8.15	0.83

Table 11: Additional statistics under rubric/table shifts: SRCC and prediction distribution (mean/std) for each judge across different anchor tables.

Metrics beyond SRCC. For each judge and each anchor table, we report: (1) SRCC, measuring whether relative ordering is preserved, and (2) the mean and standard deviation of predicted scores over the evaluation set, characterizing how the *score distribution* responds to the rubric. These statistics are summarized in Table 11.

Interpretation. A robust rubric-aware judge should ideally maintain similar SRCC across anchor tables (ranking stability), while its prediction distribution (*mean/std*) should change consistently with the table's intended shift (e.g., wider tables inducing larger dispersion, upward-shifted tables increasing the mean). As shown in Table 11, our method best matches this desideratum: it keeps ranking behavior stable while adjusting the scale/location of numeric outputs in accordance with the anchor table, indicating that it has learned to decouple comparative placement from dataset-specific numeric conventions.

C.3 Variance Across Runs

To assess robustness, we run Stage 2 evaluation five times (temperature = 0.8, top_p = 0.8) and report mean ± std in Table 12. The small standard deviations confirm that our results are stable.

Dataset	SRCC	PLCC
AADB	0.682 ± 0.0053	0.682 ± 0.0056
AVA	0.480 ± 0.0037	0.505 ± 0.0048
CADB	0.608 ± 0.0067	0.618 ± 0.0062
TAD66K	0.289 ± 0.0016	0.308 ± 0.0043

Table 12: Stage 2 variance across 5 runs.