PORTRAITAES: INTENT-CONDITIONED STRUCTURED PORTRAIT AESTHETIC ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Portrait aesthetic assessment assigns comparable scores according to how effectively human-centered images fulfill their photographic intent. These scores support data filtering, candidate selection, and preference modeling in image-generation pipelines. Existing methods typically predict a single aesthetic score or use general-purpose MLLMs without conditioning on photographic intent. This omission matters because the same blur, pose, lighting, or framing choice may serve one photographic intent but undermine another. These models thus learn context-agnostic aesthetic priors and yield inconsistent, inaccurate, misleading judgments for portraits with distinct photographic objectives. We introduce PortraitAes-Bench, an 11K-scale benchmark that decomposes this task into intent-conditioned subjudgments. Expert-authored rubrics define nine photographic intents, six first-level dimensions, and 22 secondary criteria. They support a structured pipeline for intent routing, specialist assessment, verification, and score fusion. Following this structure, we train PortraitAes with multi-task supervision. We then improve score comparability through Gaussian score calibration and within-dimension cross-image ranking. On the standard benchmark, PortraitAes achieves a Pearson correlation of 0.924 and a Spearman rank correlation of 0.934. On the hard-case set, its Pearson correlation is 0.829 and its Spearman rank correlation is 0.795. Across both sets, PortraitAes outperforms the evaluated general-purpose MLLMs and specialized aesthetic baselines.

1 INTRODUCTION

Large-scale image generation pipelines rely on automated aesthetic assessment for data filtering, candidate output selection, and preference or reward signal learning (Schuhmann et al., 2022; Kirstain et al., 2023; Xu et al., 2023). In this work, we investigate *portrait aesthetic assessment*: given a human-centric image, the task aims to interpret how its visual composition supports photographic intent and encode such perceptual judgments into comparable scalar scores. Unlike facial attractiveness prediction, this task evaluates the holistic presentation of a subject, encompassing framing, pose, gesture, action, appearance, and contextual interplay with the surrounding scene.

Portraits serve functional, commercial, documentary, social, and artistic purposes. The same blur, lighting, framing, or pose may weaken an identity portrait yet support a narrative or artistic one. A useful assessor must therefore infer intent, diagnose strengths and failures in structure, composition, image quality, tone, theme and style, and emotion, ground each judgment in visible evidence, and preserve ordering across portraits. Portrait assessment is thus an intent-conditioned, structured evaluation task rather than single-score prediction.

Existing resources cover generic aesthetic ratings, perceptual quality, multimodal critique, and portrait-specific attributes (Murray et al., 2012; Kong et al., 2016; Huang et al., 2024; Zhou et al., 2024; Huang et al., 2022; Sha et al., 2026). Their labels seldom expose photographic intent, dimension-specific rules, visible evidence, and intent-conditioned fusion together. Single holistic scores are incapable of pinpointing specific flaws in an image, whereas naive scoring with MLLMs suffers from the absence of consistent photographic standards and criterion-aligned optimization targets. The missing supervision is not another global preference score, but an executable account of which evidence matters under a given intent and how the resulting judgments should be combined.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce PortraitAes-Bench, an 11K-scale benchmark that operationalizes expert-authored rubrics through a router–specialist–fusion workflow. A router infers one of nine photographic intents and its dimension weights; six specialists assess structure, composition, image quality, tone, theme and style, and emotion; intent-conditioned fusion produces the overall score. Gemini 3 Flash (Google DeepMind, 2025) executes these roles under fixed rubrics, followed by schema checks, Qwen3.5 Flash screening, and expert verification. Each of 1.6K images yields one routing record and six specialist records, producing 11K records with scores, subcriteria, semantic labels, visible-evidence rationales, and veto/cap decisions. The multi-agent construction therefore mirrors the structure of the task instead of asking one model for an unconstrained holistic judgment.

Using the same seven-task structure, PortraitAes trains a shared Qwen3.5-VL-9B model (Qwen Team, 2026) through multi-task supervised fine-tuning (SFT). Gaussian rewards then anchor predictions to rubric-derived scores, while cross-group Bayesian personalized ranking (BPR) enforces relative order between portraits evaluated by the same specialist. Semantic, format, and Kullback–Leibler (KL) terms retain the structured outputs. PortraitAes achieves overall PLCC/SRCC values of 0.924/0.934 on the standard benchmark and 0.829/0.795 on the hard-case set; adding cross-group BPR to Gaussian-only training improves both hard-case correlations by 0.040.

Our main contributions are summarized as follows:

- **An intent-conditioned portrait aesthetic benchmark.** PortraitAes-Bench turns expert-authored rubrics into 11K structured records for 1.6K portraits, covering nine photographic intents, six dimensions, and 22 secondary criteria through intent routing, specialist assessment, and score fusion.
- **A calibrated and rank-aligned portrait aesthetic model.** PortraitAes learns the same routing and specialist structure through multi-task SFT. Gaussian calibration aligns predicted scores with rubric-derived references, while same-dimension cross-group ranking improves relative ordering across portraits.
- **Empirical gains on standard and hard cases.** PortraitAes achieves the strongest correlations among the evaluated general-purpose and specialized baselines on both the standard benchmark and its range-restricted hard-case set. Component ablations isolate the effects of structured SFT, Gaussian calibration, and cross-group ranking.

2 Related Work

2.1 General-purpose image aesthetic and quality assessment

Portrait assessment builds on image aesthetics and perceptual quality assessment. General aesthetic work spans population ratings and attributes, score distributions, visual backbones, native-resolution quality, and personalization (Murray et al., 2012; Kong et al., 2016; Talebi & Milanfar, 2018; Hosu et al., 2019; Ke et al., 2021; Yang et al., 2022). In-the-wild IQA instead emphasizes authentic distortions and capture conditions (Hosu et al., 2020; Fang et al., 2020; Ying et al., 2020). Multimodal benchmarks add critique, language-grounded judgments, and perceptual dimensions (Vera Nieto et al., 2022; Huang et al., 2024; Zhou et al., 2024; Cao et al., 2026; 2025; Li et al., 2026). Together they provide broad supervision but rarely expose photographic intent. PortraitAes-Bench links intent to portrait evidence and score fusion.

2.2 Portrait-specific aesthetic and quality assessment

Portrait-specific work adapts these objectives to human-centered images. Prior studies cover spatial and lighting features, beauty prediction, AVA-derived protocols, MPAD/MPAAD, smartphone quality, and quality question answering (Khan & Vogel, 2012; Redi et al., 2015; Kairanbay et al., 2016; Huang et al., 2022; Chahine et al., 2023; Wu et al., 2026). PortraitCraft studies portrait composition through global scores, attributes, visual question answering, and structured descriptions (Sha et al., 2026). PortraitAes-Bench instead evaluates full human-centered images, not facial attractiveness or composition alone. Its six dimensions connect body framing, pose, action, gesture, posture, and scene interaction to explicit photographic intent.

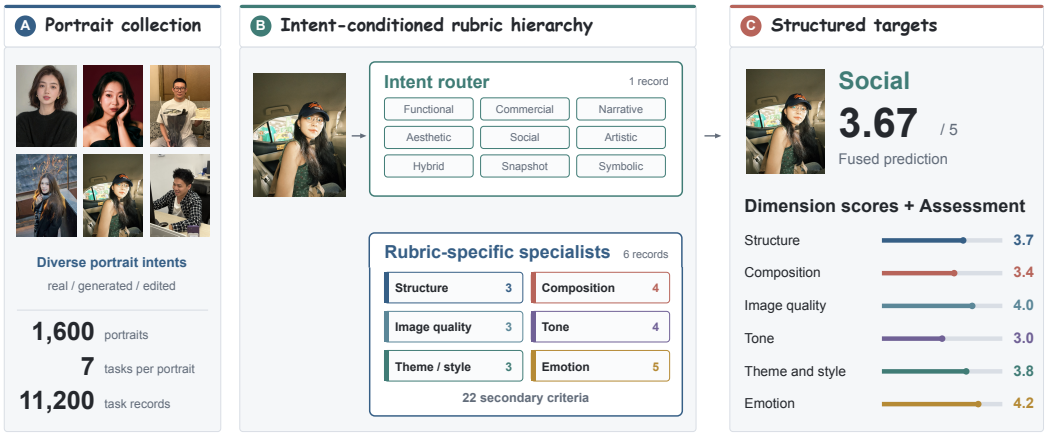

Figure 1: Overview of PortraitAes-Bench. Each portrait is routed to one of nine photographic intents and evaluated by six specialists spanning 22 secondary criteria, yielding structured routing, rating, and intent-conditioned fusion targets.

2.3 Aesthetic preference learning and reward models

Preference learning connects aesthetic assessment with reward modeling for image generation. Aesthetic supervision uses absolute ratings or score distributions, while pairwise objectives learn preference (Kong et al., 2016; Talebi & Milanfar, 2018). For text-to-image generation, Pick-a-Pic provides prompt-conditioned choices, while ImageReward and HPS v2 learn preference scorers from comparisons (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023b). MPS conditions evaluation on aesthetics, semantic alignment, detail quality, or overall preference (Zhang et al., 2024). These methods provide general rewards but do not organize portrait assessment by photographic intent and criterion-specific evidence. PortraitAes calibrates six rubric-defined dimension scores and performs ranking exclusively over examples sharing the same specialist dimension, while preserving their semantic labels and structured visual evidence.

3 PortraitAes-Bench

3.1 Benchmark overview

PortraitAes-Bench contains 1.6K portraits expanded into 11K task records, with one intent-routing record and six specialist records per image. It covers nine photographic intents, six dimensions (structure, composition, image quality, tone, theme and style, and emotion), and 22 secondary criteria. The records expose routing targets, dimension and subcriterion scores, semantic labels, visible-evidence rationales, veto/cap decisions, and an intent-conditioned overall score. Figure 1 summarizes this task structure; Appendix A compares related resources.

The 11K records represent seven linked tasks for every portrait, rather than a collection of 11K independent portrait images with unrelated labels. The routing record supplies intent and fusion weights, while each specialist record preserves a dimension-specific judgment and its evidence. This organization supports separate evaluation of routing and specialist behavior as well as end-to-end evaluation after fusion.

3.2 From photographic expertise to structured judgment

Photographic expertise specifies what to inspect and how to judge it. Motion blur, for example, must be evaluated by whether the subject remains recognizable, whether the blur supports the depicted action, and whether that choice fits a functional, narrative, or artistic intent. PortraitAes-Bench encodes this reasoning through intent categories, specialist criteria, score anchors, and decisions.

The design separates intent routing, specialist assessment, and score fusion. Routing establishes the photographic context and selects importance weights; specialists evaluate evidence under dimension-specific rubrics; fusion combines their judgments according to intent. This separates how well a criterion is realized from how much it matters to the portrait's function. A portrait can therefore be technically clean but emotionally weak, compositionally unconventional but artistically coherent, or noisy yet appropriate for a documentary or snapshot intent.

For image x, rubric $\mathcal{R}$ specifies the task boundary, labels, scoring criteria, output schema, and veto/cap rules. The router predicts intent c, its six-dimensional weight vector $w^{(c)}$, and visible routing evidence e_{type}:

$$(c, w^{(c)}, e_{type}) = f_{type}(x; \mathcal{R}_{type}). \tag{1}$$

Appendix Table 6 reports the weight matrix. Let $\mathcal{D}_{spec}$ denote the six specialist dimensions. For each $d \in \mathcal{D}_{spec}$, rubric $\mathcal{R}_d$ defines its labels, subcriteria, scoring rules, and output fields. Given the image and router context, the specialist predicts

$$z_d = (s_d, a_d, q_d, e_d, v_d) = f_d(x, c; \mathcal{R}_d), \tag{2}$$

where $s_d \in [1, 5]$ is the dimension score, a_d denotes semantic labels, q_d the subcriterion scores, e_d the evidence-grounded rationale, and v_d veto/cap decisions. For available specialists $D(x)$, the fused score is

$$S(x) = \frac{\sum_{d \in D(x)} w_d^{(c)} s_d(x)}{\sum_{d \in D(x)} w_d^{(c)}}. \tag{3}$$

Unavailable dimensions are removed from both sums rather than treated as zeros. Using reference or predicted routing weights also separates routing errors from specialist scoring errors.

3.3 Executing expert rubrics for annotation

The annotation pipeline turns these definitions into records through rubric-bound roles and quality control. Experts define the nine intents, six dimensions, 22 secondary criteria, score anchors, label spaces, task boundaries, and veto/cap rules. For each image, Gemini 3 Flash (Google DeepMind, 2025) executes the intent router and six specialist roles under their corresponding rubrics. Each role produces a separate structured record, connecting expert-defined criteria and exceptions to visible portrait evidence rather than relying on one unconstrained score. The schemas retain scores, labels, rationales, and rule decisions for supervision and audit.

Execution follows the dependency in the task definition. The router first produces the intent, weights, and routing evidence. Its context is then supplied to all six specialists, which inspect the same portrait through different criteria and return independently parseable records. Rule-based fusion is applied only after these records pass quality control, so the intermediate judgments remain available instead of being hidden inside the final score.

The rubrics specify contextual priorities and concrete observations. Functional and commercial intents prioritize recognizability and finish; narrative and symbolic intents emphasize theme and emotion; aesthetic and artistic intents allow stylization; social and snapshot intents emphasize presentation and affect. Hybrid intent balances these priorities. Structure covers subject prominence, completeness, and background harmony; composition covers framing, perspective, leading lines, and dynamism. Image quality examines sharpness, texture, noise, compression, and intent-sensitive tolerance. Tone relates light–dark organization and color treatment to brightness, purity, and harmony. Theme and style evaluates fit, consistency, and originality, while emotion connects affect and intensity to physiological cues, context, and body language.

Before fusion, schema checks enforce the task contract and valid fields. Qwen3.5 Flash (Alibaba Cloud, 2026) screens format, score ranges, label compatibility, and veto/cap consistency; experts verify rubric consistency and boundary cases. Specialist-level veto/cap decisions prevent hard dimension failures from being averaged away. Figure 2 summarizes the process.

3.4 Artifact statistics

After parsing, consistency screening, fusion checks, and expert verification, the artifact contains 1.6K images and 11K task records. Each image contributes one routing and six specialist records,

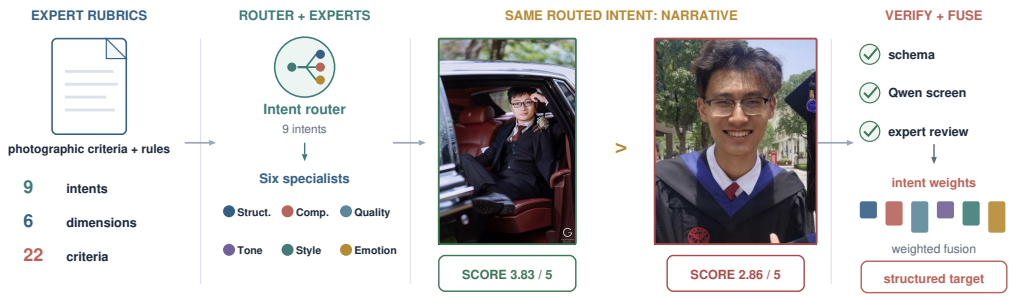

Figure 2: Construction of PortraitAes-Bench. Expert rubrics define an intent router and six specialist roles; schema checks, Qwen3.5 Flash screening, and expert review precede intent-weighted fusion.

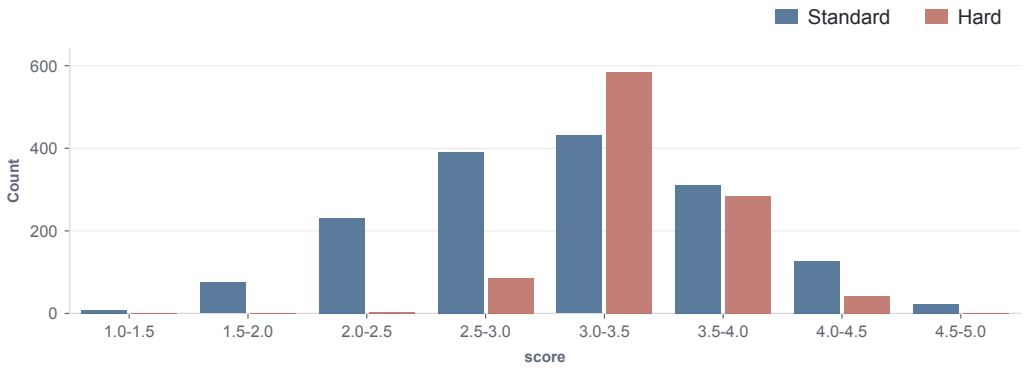

Figure 3: Intent-conditioned overall-score distributions. The standard benchmark spans a broad range, whereas the hard-case set concentrates visually similar portraits for fine-grained ranking.

supporting image-level and task-specific evaluation of routing, dimension-wise regression, semantic recognition, evidence-grounded explanation, and intent-conditioned scoring.

Appendix B summarizes the specialist-score distributions (Figure 5). Emotion and structure concentrate higher, tone is lower and broader, and the remaining dimensions occupy the middle range, providing varied cases for calibration. The overall distributions contrast the broad standard benchmark with the deliberately range-restricted hard-case set (Figure 3).

4 PortraitAes: Intent-Conditioned Portrait Aesthetic Modeling

A portrait aesthetic model must explain its scores and compare images consistently. Holistic supervision leaves intent and individual criteria implicit; structured supervision exposes them, but independent predictions still do not constrain cross-image order and can rank a weaker portrait above a stronger one under the same criterion. Multi-task SFT first learns intent and specialist judgments under shared rubrics. Reinforcement learning (RL) then combines Gaussian calibration with same-dimension cross-group ranking, while semantic and format rewards preserve structured outputs (Figure 4).

4.1 Learning structured aesthetic priors

Each training image yields one intent-routing target and six specialist targets, with a task-specific rubric and output schema. We construct the instruction–response examples using the same annotation protocol as PortraitAes-Bench. The SFT corpus contains 10,286 images and approximately 72,000 examples across seven tasks, with 10,286 examples for each task: intent routing, theme and style, composition, structure, image quality, tone, and emotion. Qwen3.5 Flash normalizes output

formats and corrects schema inconsistencies, followed by manual spot-checking for format compliance and annotation quality.

Let $a \in \mathcal{A} = \{r\} \cup \mathcal{D}$ index the routing task r or one of the six specialist dimensions $d \in \mathcal{D}$, and let q_a denote its task-specific instruction. Given an image I and a structured target response y, we optimize the autoregressive objective

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(I,q_a,y)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\sum_{t=1}^{|y|}\log \pi_\theta(y_t \mid I, q_a, y_{<t})\right]. \tag{4}$$

Task conditioning lets one policy learn heterogeneous portrait rubrics over a shared visual representation without collapsing their label spaces or schemas. The same model can therefore execute one specialist in isolation or the complete router–specialist workflow. We apply low-rank adaptation (LoRA) to the Qwen3.5-VL-9B language backbone and freeze its visual encoder and aligner, concentrating adaptation on rubric interpretation, structured prediction, and score generation.

4.2 Rank-aligned aesthetic optimization

Token-level SFT learns rubric semantics and response structure but neither weights rating errors by numerical distance nor compares image queries. Predicting 1 for a reference of 5 is numerically worse than predicting 4, yet cross-entropy depends on target-token probability. Scores must therefore agree with their rubric-derived references and preserve same-dimension ordering. Small pointwise errors can reverse nearby scores, while predictions concentrated near the scale midpoint may have modest error but weak discrimination. We combine calibration with comparisons across prompt groups to address both problems.

4.3 Intent-conditioned reward design

The completion reward combines calibration, ranking, semantic consistency, and structural validity:

$$R_i = \lambda_{\mathrm{cal}}R_{\mathrm{cal},i} + \lambda_{\mathrm{rank}}R_{\mathrm{rank},i} + \lambda_{\mathrm{sem}}R_{\mathrm{sem},i} + \lambda_{\mathrm{fmt}}R_{\mathrm{fmt},i}, \qquad \lambda_\bullet \geq 0. \tag{5}$$

For specialist tasks, R_{cal} maps primary-score and subcriterion errors to Gaussian soft rewards (Cao et al., 2025). R_{rank} supplies the missing comparison signal: each rollout is compared only with valid cross-group examples from the same specialist dimension, using a BPR-style reward (Rendle et al., 2012) when the reference gap defines a reliable order. This constraint prevents comparisons between incompatible rubrics such as structure and emotion.

Absolute calibration (R_{cal}). Inspired by Gaussian soft rewards for visual rating (Cao et al., 2025), we map prediction error to a smooth calibration signal. For a scalar prediction p and its rubric-derived reference g, we define

$$G(p, g) = \exp\left[-\frac{(p-g)^2}{2\sigma^2}\right], \tag{6}$$

where σ controls the tolerance to score deviation. For a specialist dimension d with primary score (p_d, g_d) and rubric subcriteria $\mathcal{S}_d$, the calibration reward is

$$R_{\mathrm{cal}}^{(d)} = \eta_d G(p_d, g_d) + (1-\eta_d)\sum_{k\in\mathcal{S}_d}\omega_{d,k}G(p_{d,k}, g_{d,k}), \qquad \sum_{k\in\mathcal{S}_d}\omega_{d,k} = 1. \tag{7}$$

A specialist response must calibrate both its overall judgment and the evidence-bearing sub-scores.

Cross-group relative ranking (R_{rank}). Calibration evaluates each image independently and cannot constrain order between images. We therefore pool predictions from different GRPO completion groups within an update, which generally correspond to different image prompts. Each rollout is compared with valid same-dimension rollouts from other groups, and its ranking reward aggregates the pairwise terms.

Let $\mathcal{G}_i$ denote the prompt group that produced response i. For each specialist response, we construct the cross-group comparator set

$$\mathcal{P}_d(i) = \{j :\ \mathcal{G}_j \neq \mathcal{G}_i,\ d_j = d_i = d,\ |g_i - g_j| \geq \delta_d\}. \tag{8}$$

The minimum reference gap δ_d removes ambiguous pairs. For each valid pair, a centered BPR-inspired reward (Rendle et al., 2012) is

$$r_{ij}^{(d)} = 2\,\sigma_L\left(\frac{\mathrm{sign}(g_i - g_j)(p_i - p_j) - m_d}{\tau_d}\right) - 1, \tag{9}$$

where $\sigma_L(z) = (1 + e^{-z})^{-1}$, m_d is the ordering margin, and τ_d controls smoothness. The reward is positive for a correct predicted order and negative for a reversal. We aggregate pair rewards using reference-score separation as a confidence weight:

$$R_{\mathrm{rank},i}^{(d)} = \begin{cases} \dfrac{\sum_{j\in\mathcal{P}_d(i)} |g_i - g_j|\, r_{ij}^{(d)}}{\sum_{j\in\mathcal{P}_d(i)} |g_i - g_j|}, & \mathcal{P}_d(i) \neq \emptyset, \\ 0, & \mathcal{P}_d(i) = \emptyset. \end{cases} \tag{10}$$

The reference gap selects and weights reliable comparisons rather than serving as a regression target; router records are excluded. Calibration constrains absolute scores, whereas ranking constrains cross-image order. Their combination avoids trading one objective for the other: ranking need not move all predictions away from the rubric-derived scale, and calibration alone need not determine the ordering of visually close portraits.

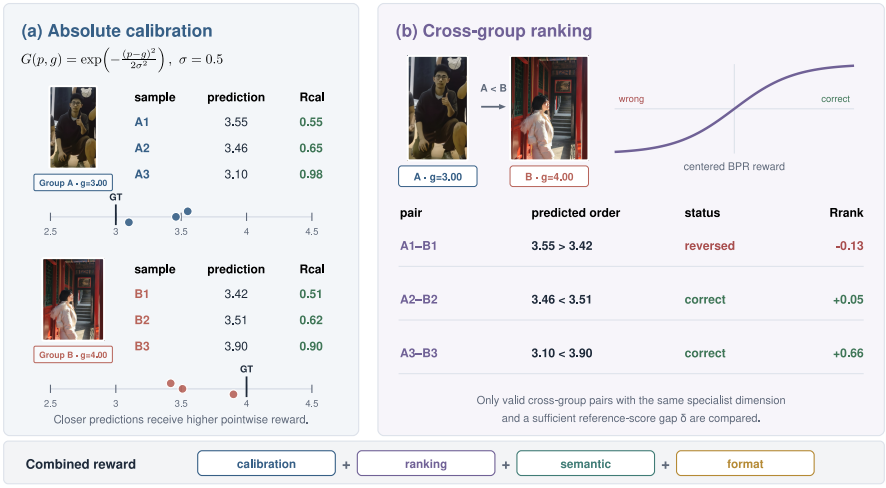

Figure 4: Illustration of the numerical rewards. Gaussian calibration anchors predictions to rubric-derived scores; same-dimension cross-group ranking penalizes ordinal inversions. Images and roll-out values illustrate the mechanism rather than report an evaluation case.

Semantic and format rewards enforce rubric-compatible labels and parseable schemas. Router records replace calibration with intent-category and fusion-weight rewards and disable ranking. Appendices D, C, and E provide the auxiliary rewards, policy objective, and RL settings, respectively.

5 Experiments

5.1 Experimental setup

We train PortraitAes from Qwen3.5-VL-9B in the two-stage procedure described in Section 4. The SFT stage uses seven task-specific instruction corpora constructed with the same rubric-driven annotation protocol as PortraitAes-Bench, and the RL stage starts from the SFT adapter and optimizes the intent-conditioned reward with Dr.GRPO (Liu et al., 2024). The visual encoder and aligner are frozen in both stages, and LoRA is applied to the language backbone. Full optimization settings, reward weights, and compute details are provided in Appendix E.

5.2 Evaluation protocol

The standard PortraitAes-Bench evaluation contains 1.6K images expanded into 11K task-specific examples, and the hard-case evaluation contains 1K held-out images expanded into 7K task exam-

Table 1: Standard PortraitAes-Bench results (PLCC/SRCC; higher is better). † denotes diagnostic correlations from a holistic-only scorer.

Model	Structure		Composition		Quality		Tone		Theme		Emotion		Overall	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Qwen3.5-VL-9B	0.833	0.822	0.724	0.742	0.741	0.698	0.807	0.789	0.706	0.714	0.698	0.740	0.857	0.874
InternVL3-8B	0.393	0.453	0.518	0.505	0.581	0.623	0.530	0.570	0.448	0.511	0.297	0.296	0.607	0.668
PortraitCraft-4B†	0.698	0.801	0.647	0.760	0.738	0.773	0.578	0.698	0.613	0.754	0.535	0.655	0.652	0.785
Qwen3.5 Flash	0.791	0.837	0.620	0.700	0.807	0.798	0.795	0.818	0.590	0.705	0.667	0.723	0.804	0.763
ArtiMuse†	0.741	0.785	0.744	0.796	0.556	0.602	0.688	0.717	0.725	0.756	0.712	0.703	0.785	0.804
UniPercept†	0.648	0.709	0.665	0.711	0.624	0.648	0.631	0.655	0.676	0.711	0.625	0.652	0.703	0.729
HumanAesExpert-8B	0.537	0.574	0.541	0.567	0.583	0.546	0.465	0.483	0.414	0.460	0.361	0.402	0.457	0.499
Q-Align†	0.717	0.764	0.729	0.752	0.715	0.705	0.679	0.697	0.652	0.684	0.604	0.622	0.707	0.724
PortraitAes (Ours)	**0.875**	**0.912**	**0.878**	**0.903**	**0.889**	**0.891**	**0.914**	**0.913**	**0.879**	**0.901**	**0.847**	**0.863**	**0.924**	**0.934**

Table 2: PortraitAes-Bench hard-case results (PLCC/SRCC; higher is better). † denotes diagnostic correlations from a holistic-only scorer.

Model	Structure		Composition		Quality		Tone		Theme		Emotion		Overall	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Qwen3.5-VL-9B	0.515	0.513	0.346	0.281	0.558	0.506	0.602	0.574	0.385	0.474	0.391	0.413	0.649	0.629
InternVL3-8B	0.030	0.036	0.156	0.112	0.402	0.476	0.228	0.273	-0.010	-0.033	0.121	0.184	0.210	0.215
PortraitCraft-4B†	0.406	0.403	0.361	0.331	0.479	0.463	0.151	0.145	0.342	0.328	0.288	0.297	0.406	0.400
Qwen3.5 Flash	0.522	0.522	0.240	0.214	0.578	0.543	0.624	0.601	0.304	0.411	0.407	0.400	0.615	0.635
ArtiMuse†	0.430	0.421	0.407	0.398	0.336	0.346	0.368	0.350	0.410	0.418	0.387	0.383	0.501	0.495
UniPercept†	0.292	0.319	0.296	0.271	0.322	0.349	0.151	0.146	0.322	0.331	0.296	0.294	0.352	0.358
HumanAesExpert-8B	0.243	0.237	0.197	0.194	0.200	0.178	0.110	0.105	0.106	0.077	0.100	0.094	0.127	0.154
Q-Align†	0.422	0.400	0.360	0.339	0.451	0.411	0.223	0.215	0.277	0.242	0.229	0.227	0.374	0.348
PortraitAes (Ours)	**0.679**	**0.650**	**0.646**	**0.614**	**0.768**	**0.750**	**0.789**	**0.752**	**0.689**	**0.627**	**0.650**	**0.629**	**0.829**	**0.795**

ples. We report Pearson's linear correlation coefficient (PLCC) for score calibration and Spearman's rank correlation coefficient (SRCC) for ordinal consistency. Comparisons cover general-purpose MLLMs, the open InternVL3-8B checkpoint (Zhu et al., 2025), and specialized aesthetic or perceptual assessment systems, including PortraitCraft (Sha et al., 2026; Bai et al., 2025), ArtiMuse (Cao et al., 2026), UniPercept (Cao et al., 2025), HumanAesExpert-8B (Liao et al., 2025), and Q-Align (Wu et al., 2023a). PortraitAes overall scores are fused using predicted intent-routing weights, whereas baselines without an intent router use their native holistic outputs. For holistic-only scorers, daggered dimension entries report the correlation between their single holistic prediction and each PortraitAes-Bench target as a diagnostic rather than a dimension-conditioned estimate. Additional baseline and split details are given in Appendix E.

5.3 Benchmark results and model analysis

PortraitAes leads every standard-set metric, reaching mean specialist PLCC/SRCC values of 0.880/0.897 and overall values of 0.924/0.934 (Table 1). Its largest gains are on theme and style and emotion, which require intent-aware judgments of pose, scene, styling, and affect.

General-purpose MLLMs remain uneven, while holistic systems such as ArtiMuse, UniPercept, and Q-Align provide only diagnostic covariance with dimension targets. PortraitCraft-4B transfers better to structure and composition but is weaker on tone, theme and style, and emotion. By predicting intent and specialist scores before fusion, PortraitAes makes the final estimate traceable to benchmark criteria.

5.4 HARD-CASE ANALYSIS

The hard-case set narrows the overall-score standard deviation from 0.688 to 0.315 and the mean specialist deviation from 0.913 to 0.457 (panel (b) of Figure 3). In this compressed range, PortraitAes retains mean specialist PLCC/SRCC values of 0.704/0.670 and overall values of 0.829/0.795 (Table 2). Its margin over ArtiMuse (Cao et al., 2026) grows from 0.139/0.130 on the standard benchmark to 0.328/0.300 on the hard set, highlighting the strength of calibrated ranking on visually similar portraits.

Cross-group BPR is especially helpful for structure and composition, where small framing or subject-presentation differences can reverse order without producing large score gaps. Gaussian calibration anchors the score scale, whereas ranking directly penalizes same-dimension ordinal inversions (Table 4).

5.5 COMPONENT ABLATIONS

The ablations isolate the effect of structured SFT, Gaussian calibration, and cross-group BPR before intent-conditioned fusion.

Table 3: Component ablation on the standard PortraitAes-Bench. Each result is reported as PLCC/SRCC; higher is better. Checkmarks denote enabled training components.

Method	Training component			PLCC/SRCC						
	SFT	Gaussian	Cross-BPR	Structure	Composition	Quality	Tone	Theme	Emotion	Overall
Qwen3.5-VL-9B				0.833/0.822	0.724/0.742	0.741/0.698	0.807/0.789	0.706/0.714	0.698/0.740	0.857/0.874
PortraitAes-SFT	✓			0.856/0.896	0.852/0.885	0.873/0.868	0.904/0.905	0.855/0.877	0.810/0.830	0.909/0.917
PortraitAes-GRPO (Gaussian only)	✓	✓		0.863/0.901	0.869/0.897	0.872/0.868	0.907/0.907	0.854/0.883	0.820/0.837	0.915/0.924
PortraitAes-GRPO	✓	✓	✓	**0.875/0.912**	**0.878/0.903**	**0.889/0.891**	**0.914/0.913**	**0.879/0.901**	**0.847/0.863**	**0.924/0.934**

Table 4: Component ablation on the PortraitAes-Bench hard-case set. Each result is reported as PLCC/SRCC; higher is better. Checkmarks denote enabled training components.

Method	Training component			PLCC/SRCC						
	SFT	Gaussian	Cross-BPR	Structure	Composition	Quality	Tone	Theme	Emotion	Overall
Qwen3.5-VL-9B				0.515/0.513	0.346/0.281	0.558/0.506	0.602/0.574	0.385/0.474	0.391/0.413	0.649/0.629
PortraitAes-SFT	✓			0.562/**0.668**	0.558/0.524	0.737/0.723	**0.795/0.753**	0.570/0.598	0.549/0.559	0.780/0.758
PortraitAes-GRPO (Gaussian only)	✓	✓		0.617/0.624	0.597/0.540	0.713/0.694	0.768/0.731	0.661/0.590	0.610/0.587	0.789/0.755
PortraitAes-GRPO	✓	✓	✓	**0.679**/0.650	**0.646/0.614**	**0.768/0.750**	0.789/0.752	**0.689/0.627**	**0.650/0.629**	**0.829/0.795**

On the standard set, structured SFT raises overall PLCC/SRCC from 0.857/0.874 to 0.909/0.917, Gaussian calibration reaches 0.915/0.924, and cross-group BPR reaches 0.924/0.934 (Table 3). On the hard set, BPR raises the Gaussian-only result from 0.789/0.755 to 0.829/0.795 and improves all six dimensions. Calibration therefore stabilizes absolute scores, while cross-group ranking preserves fine-grained order in a compressed score range.

6 CONCLUSION

We introduced PortraitAes-Bench, an 11K-scale benchmark that represents portrait aesthetics as intent-conditioned judgments across nine photographic intents and six dimensions. Its router–specialist–fusion workflow turns expert rubrics into structured supervision with scores, labels, visual evidence, and rule decisions. PortraitAes learns the same task structure through multi-task SFT, then combines Gaussian calibration with same-dimension cross-image ranking. Results on the standard and hard-case sets show that this combination improves both agreement with rubric-derived scores and ordinal consistency over the evaluated general-purpose and specialized baselines.

REFERENCES

Alibaba Cloud. qwen3.5-flash model info. Alibaba Cloud Model Studio documentation, 2026. URL https://help.aliyun.com/en/model-studio/qwen3-5-flash.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, et al. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Theo Cayla, Sira Ferradans, and Jean Ponce. An image quality assessment dataset for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9968–9978, 2023.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683. IEEE, 2020.

Google DeepMind. Gemini 3 Flash model card. Google DeepMind model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9375–9383, 2019.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Heng Huang, Xin Jin, Xinning Li, Shuai Cui, and Chaoen Xiao. Aesthetic evaluation of asian and caucasian photos with overall and attribute scores. *Computers and Electrical Engineering*, 103: 108341, 2022.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024.

Magzhan Kairanbay, John See, and Lai-Kuan Wong. Aesthetic evaluation of facial portraits using compositional augmentation for deep cnns. In *Asian conference on computer vision*, pp. 462–474. Springer, 2016.

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

Shehroz S Khan and Daniel Vogel. Evaluating visual aesthetics in photographic portraiture. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, visualization, and imaging*, pp. 55–62, 2012.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

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

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026.

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

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. *URL https://arxiv.org/abs/2503.20783*, 1, 2024.

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

Qwen Team. Qwen3.5-9B model card. Hugging Face model repository, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`.

Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. The beauty of capturing faces: Rating the quality of digital portraits. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pp. 1–8. IEEE, 2015.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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.

Yuyang Sha, Zijie Lou, Youyun Tang, Xiaochao Qu, Zheng Qu, Ben Xia, Haoxiang Li, Ting Liu, and Luoqi Liu. Portraitcraft: A benchmark for portrait composition understanding and generation. *arXiv preprint arXiv:2604.03611*, 2026.

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

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Daniel Vera Nieto, Luigi Celona, and Clara Fernandez Labrador. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. *Advances in Neural Information Processing Systems*, 35:34148–34161, 2022.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023a.

Sijing Wu, Yunhao Li, Zicheng Zhang, Qi Jia, Xinyue Li, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. Q-bench-portrait: Benchmarking multimodal large language models on portrait image quality perception. *arXiv preprint arXiv:2601.18346*, 2026.

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

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

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19829–19837. IEEE, 2022.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3572–3582. IEEE, 2020.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027. IEEE, 2024.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

This appendix provides compact supporting material for interpreting the benchmark scope, diagnostic figures, ablation results, and implementation settings.

A RESOURCE COMPARISON

Table 5 uses five coarse properties (scale, portrait focus, supervision type, dimensional structure, and textual annotation) to clarify how the benchmark differs from prior resources.

Table 5: Comparison with representative aesthetics and portrait-related resources. PortraitAes-Bench emphasizes intent-conditioned, multi-dimensional portrait understanding rather than a single global score.

Resource	Scale	Port.	Supervision	Dim.	Text
AVA (Murray et al., 2012)	255K	No	ratings	global	limited
AADB (Kong et al., 2016)	10K	No	score/attr./rank	attr.	no
ArtiMuse (Cao et al., 2026)	10K	No	score/attr./text	8 attr.	yes
UniPercept (Cao et al., 2025)	800K	No	public VR/VQA	perception	yes
Qwen-Image-Bench (Li et al., 2026)	1K	No	judge scores	5D/56F	yes
MPAD/MPAAD (Huang et al., 2022)	18K/3K	Yes	overall/attr.	3 attr.	no
PIQ23 (Chahine et al., 2023)	5K	Yes	pairwise	quality	no
Q-Bench-Portrait (Wu et al., 2026)	2.8K	Yes	QA	quality	yes
PortraitCraft (Sha et al., 2026)	50K	Yes	score/attr./VQA	comp.	yes
PortraitAes-Bench	11K	Yes	intent/6 scores/labels	6D	yes

B DIAGNOSTIC VISUALIZATIONS

This diagnostic reports how score mass varies across specialist dimensions.

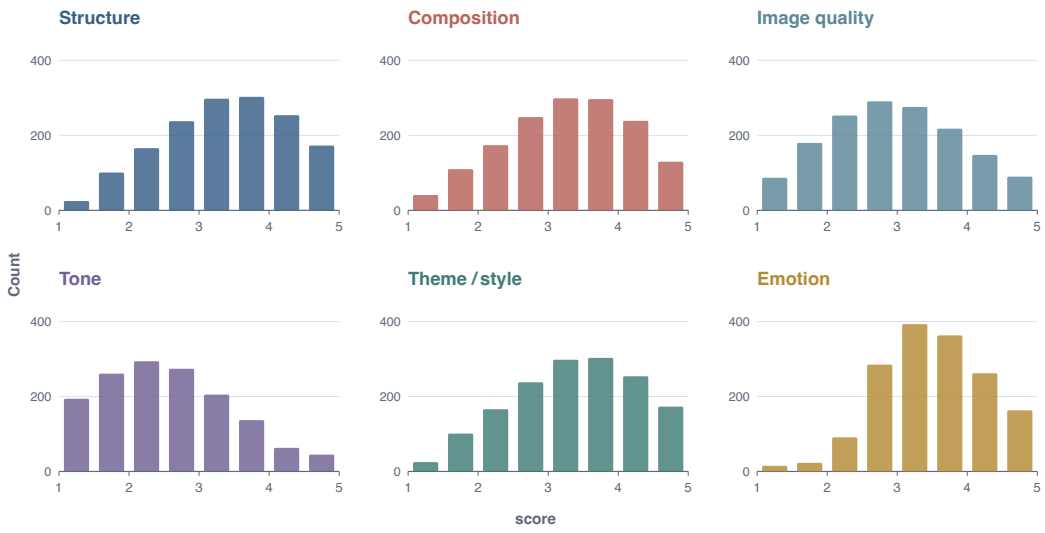

Figure 5: Specialist-score distributions in PortraitAes-Bench. The six portrait-aesthetic dimensions show distinct score profiles rather than a single shared difficulty pattern, supporting dimension-wise evaluation of structure, composition, image quality, tone, theme/style, and emotion.

C Policy Optimization Details

For an image–instruction pair $x = (I, q_a)$, let $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ be the i-th of K sampled structured responses. Following group relative policy optimization (GRPO) (Shao et al., 2024), we center rewards within each prompt group and scale them by the dispersion of the rollout batch,

$$\widehat{A}_i = \frac{R_i - \mu_{\mathcal{G}(x)}}{s_{\mathcal{B}} + \epsilon_A}, \tag{11}$$

where $\mathcal{G}(x)$ is the completion group for x and $\mathcal{B}$ is the rollout batch. The policy is optimized with a clipped surrogate objective regularized toward the SFT reference policy π_{ref}:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{Z}\sum_{t=1}^{|y_i|}\left(\min\left(\rho_{i,t}\widehat{A}_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_i\right) - \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})_{i,t}\right)\right], \tag{12}$$

where $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$, ϵ is the clipping radius, and Z is a fixed response-length normalizer. The clipped ratio limits destructive policy updates, while the token-level KL term controls drift from the schema-aware SFT model.

D AUXILIARY REWARD DEFINITIONS

Intent routing (R_{route}). The router has no scalar aesthetic target. Its reward instead combines intent-category accuracy with the closeness of the predicted fusion-weight vector $\widehat{\mathbf{w}}$ to its reference $\mathbf{w}$:

$$R_{\text{route}} = \eta_r \mathbf{1}[\widehat{c} = c] + (1 - \eta_r) \left[1 - \tfrac{1}{2}\|\widehat{\mathbf{w}} - \mathbf{w}\|_1\right]_+ . \tag{13}$$

Semantic consistency (R_{sem}). Numeric agreement is insufficient when a response assigns an incompatible semantic label. Let $s_d(y_i)$ and $s_d^\star$ be the task-specific semantic field extracted from the response and the reference. We use a compact task-specific matcher:

$$R_{\text{sem}}(y_i, d) = \rho_d(s_d(y_i), s_d^\star) , \tag{14}$$

where ρ_d assigns full credit to exact semantic matches, partial credit to compatible labels, a penalty to contradictions, a parse-failure penalty to malformed structured responses, and a neutral value when no semantic target is defined. For example, in theme and style, two subject codes with the same major code, such as $S3.1$ and $S3.4$, are compatible and receive partial credit; emotion uses nonempty category-token overlap as the compatibility check.

Structural validity (R_{fmt}). The format reward grades dimension-specific schema completeness and subtracts a length-dependent penalty,

$$R_{\text{fmt}}(y, d) = \text{clip}(B_d(y) - \ell_d(|y|),\, r_{\min}, r_{\max}) , \tag{15}$$

where B_d checks well-formed structured responses and required fields, and ℓ_d discourages unnecessarily long responses. Malformed outputs receive a parse-failure penalty and are excluded from ordinal comparison.

E Additional Experimental Details

Training setup. PortraitAes is initialized from Qwen3.5-VL-9B. For supervised fine-tuning, we apply LoRA with rank 64 and scaling factor 128 to all linear modules while freezing the visual encoder and aligner. SFT is performed in bfloat16 for three epochs using a cosine learning-rate schedule with a peak learning rate of 4×10^{-5} and 5% warmup. The maximum sequence length is 16,384, the per-device batch size is 1, and gradient accumulation is 2. Gradient checkpointing and ZeRO-3 offload are enabled, and training uses 24 NVIDIA L20Z GPUs.

For reinforcement learning, the corpus contains approximately 2,657 images expanded into 18,600 task-specific examples. Both the trainable policy and reference policy are initialized from the seven-task SFT adapter, with the visual encoder and aligner remaining frozen. We optimize the intent-conditioned reward in Section 4.3 using Dr.GRPO with batch-level reward scaling. Each prompt samples eight completions, and training runs for one epoch with a constant learning rate of 2×10^{-6} and KL coefficient 0.02, using bfloat16 precision and ZeRO-3 across 24 NVIDIA L20Z GPUs.

Evaluation details. The standard evaluation uses one routing task and six specialist rating tasks for each of 1,600 images. The hard-case set is image-disjoint and contains 1,000 images from a held-out portrait pool after exact- and basename-level exclusion against training and earlier evaluation artifacts. General-purpose MLLMs are evaluated with the corresponding PortraitAes-Bench rubric prompt; Qwen3.5 Flash is queried through Alibaba Cloud Model Studio using the fixed API identifier `qwen3.5-flash` (Alibaba Cloud, 2026). Specialized aesthetic models are evaluated with their released checkpoints and native scoring interfaces. All table entries are PLCC/SRCC, and unavailable evaluations are marked with "–".

Intent-conditioned fusion. Table 6 reports the router weights used to convert six specialist scores into an overall portrait-aesthetic score. Functional and commercial portraits emphasize quality and presentation, aesthetic and social portraits rely more on tone and theme and style, and snapshot or symbolic portraits shift toward emotion and meaning.

Table 6: Intent-conditioned fusion weights used by the router.

Intent	Struct.	Comp.	Quality	Tone	Theme	Emotion
I Functional	0.20	0.20	0.30	0.15	0.10	0.05
II Commercial	0.10	0.15	0.35	0.25	0.10	0.05
III Narrative	0.15	0.15	0.15	0.20	0.10	0.25
IV Aesthetic	0.10	0.15	0.15	0.30	0.20	0.10
V Social	0.05	0.15	0.15	0.20	0.30	0.15
VI Artistic	0.10	0.15	0.05	0.20	0.20	0.30
VII Hybrid	0.10	0.15	0.15	0.20	0.20	0.20
VIII Snapshot	0.05	0.10	0.05	0.15	0.15	0.50
IX Symbolic	0.00	0.05	0.00	0.05	0.30	0.60

Rank-aligned GRPO. Table 7 lists the RL-stage settings. Calibration and cross-image ranking dominate the reward, while semantic and format terms remain lightweight constraints.

Table 7: Hyperparameters for rank-aligned GRPO.

Hyperparameter	Value	Hyperparameter	Value
Completions per prompt K	8	Training epochs	1
Learning rate	2×10^{-6}	Warmup ratio	0.05
KL coefficient	0.02	Clip radius	0.12
Per-device batch size	4	Gradient accumulation	1
Calibration/ranking weights	1.4 / 0.8	Semantic/format weights	0.05 / 0.02
Gaussian bandwidth σ	0.5	Primary/sub-score weights	0.8 / 0.2
Minimum reference gap δ	1.0	BPR margin / temperature	0.0 / 0.5
Sampling temperature / top-p	1.0 / 0.95	Maximum lengths	512 / 16,384
Maximum gradient norm	0.5	Precision	bfloat16

F EXAMPLES



Figure 6: Structured predictions for twelve portraits across six intent categories, grouped by category and ordered by decreasing overall score. Colored bars show the six specialist scores.

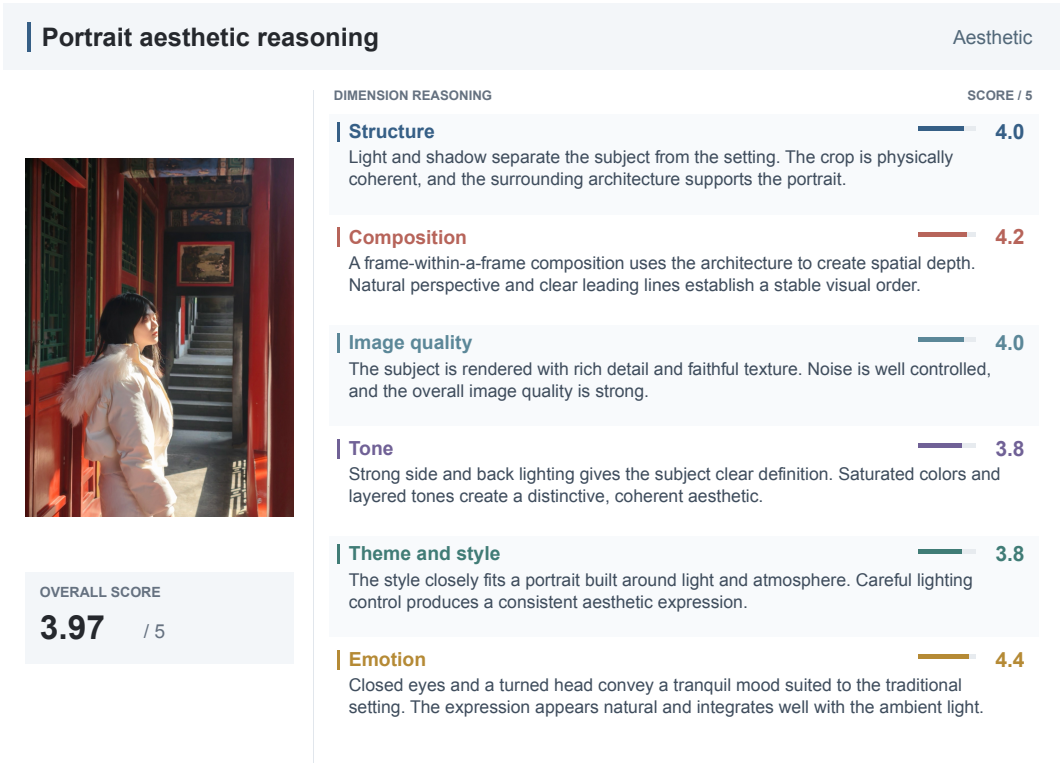

Figure 7: Dimension-level aesthetic explanations for an aesthetic portrait with an overall score of 3.97.

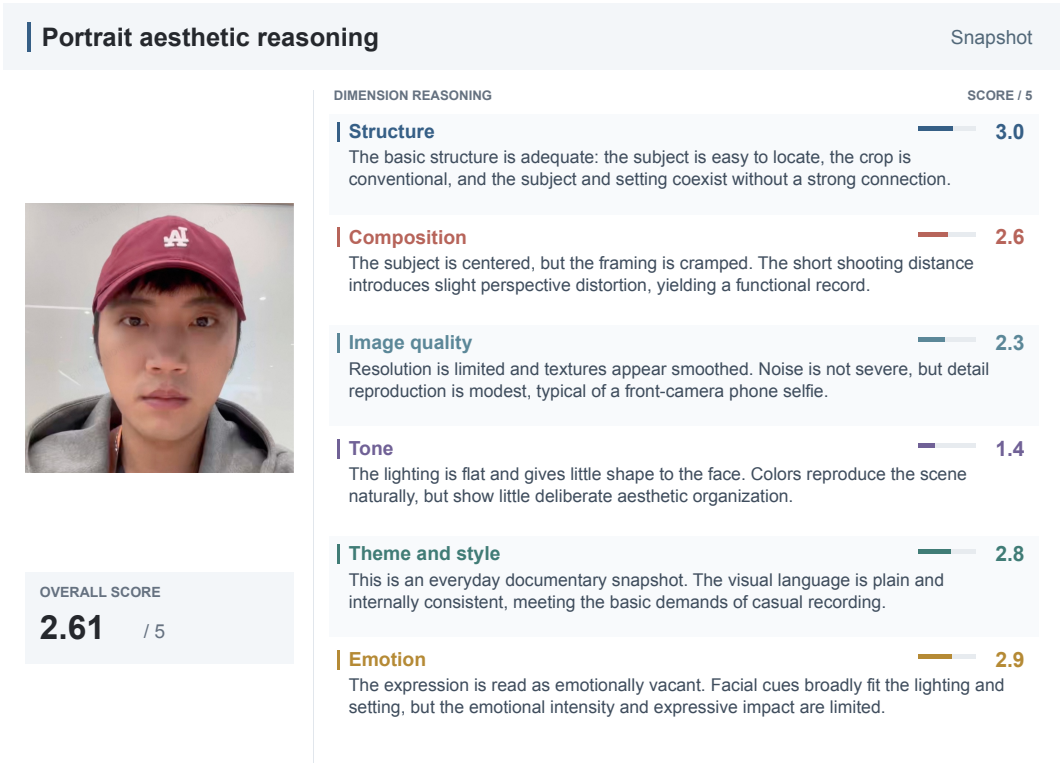

Figure 8: Dimension-level aesthetic explanations for a snapshot with an overall score of 2.61.

G LIMITATIONS

PortraitAes-Bench uses a fixed portrait-intent taxonomy and rubric-bound annotations. Broader human validation and coverage of additional photographic styles and cultural contexts remain directions for future work.

H AI USE STATEMENT

Generative AI tools assisted with conceptual organization; portions of the annotation, training, and evaluation pipeline; application of author-specified visual-language rubrics; and editorial tasks including condensation, copy-editing, LaTeX formatting, and a non-submission Chinese working translation. They did not determine the mathematical claims, experimental hypotheses, evaluation design, or empirical interpretation. The authors reviewed the AI-assisted code, annotations, manuscript edits, and reported claims, and take responsibility for the paper and associated artifacts.