ALIGNING UNDERSTANDING AND SCORING FOR DATA-EFFICIENT PHOTOGRAPHIC AESTHETIC ASSESSMENT

Anonymous authors
Paper under double-blind review

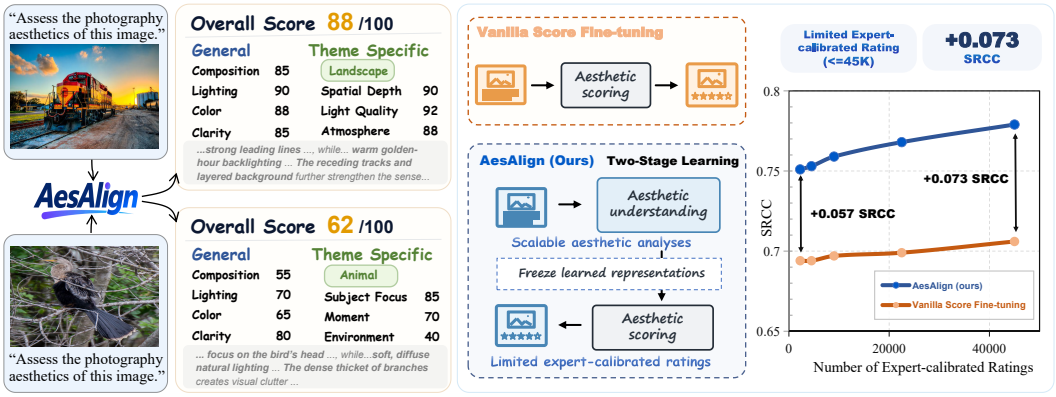


Figure 1: AesAlign provides overall and dimension-level aesthetic analyses and scores for photographs. With the same limited expert-rating supervision, it achieves higher SRCC than vanilla score fine-tuning.

ABSTRACT

In this work, we study Photography Aesthetics Assessment (PAA), which aims to provide comprehensive aesthetic evaluation of photographs through both qualitative understanding and quantitative scoring. However, reliable photography aesthetic scoring remains challenging because large-scale aesthetic analyses can be obtained relatively easily, while high-quality and consistently calibrated human ratings are costly and difficult to scale. To address these challenges, we propose **AesAlign**, a data-efficient framework that decouples photography aesthetic understanding from quantitative score calibration. First, we establish a unified photography aesthetic evaluation protocol and construct large-scale fine-grained aesthetic analyses for 500K smartphone photographs. We further collect high-quality human-verified overall and dimension-level ratings for 50K images to provide reliable quantitative supervision. Based on these complementary supervision signals, AesAlign learns photography aesthetic knowledge from scalable analyses and subsequently calibrates the learned representations to human rating scales using a lightweight scoring probe. Extensive experiments demonstrate that **AesAlign** achieves strong performance under limited human supervision and exhibits robust cross-dataset generalization, enabling comprehensive and interpretable photography aesthetics assessment with reliable overall and multidimensional scores. Our models, training code, and dataset will be publicly available.

1 INTRODUCTION

Image Aesthetics Assessment (IAA) has been extensively studied and underpins various downstream applications. Existing IAA methods typically treat photographs He et al. (2016), paintings Liu et al. (2025b), and AI-generated imagess Liu et al. (2025a) alike within a unified framework. Photographs,

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.

however, differ from generic images in many aspects. Beyond general attributes such as composition and color harmony, their aesthetics are closely tied to the image-capturing process, including exposure, depth of field, focus, and timing Qi et al. (2025). In this work, we focus on Photography Aesthetics Assessment (PAA), which evaluates the aesthetic quality of photographs with explicit attention to these photography-specific factors.

To capture these factors, a comprehensive PAA system should evaluate a photograph both as a whole and along multiple photographic dimensions. At both levels, it should provide a fine-grained aesthetic analysis together with a quantitative score (Figure 1). Analyses ground the assessment in visual evidence and offer interpretable guidance for downstream tasks Wei et al. (2018); Dutt et al. (2025), while scores distill them into a comparable basis for photo ranking and recommendation Murray et al. (2012); He et al. (2016).

A straightforward solution for PAA is to directly transfer existing IAA methods, which broadly fall into two categories. One line of work focuses on aesthetic understanding and produces analyses through critique generation, visual question answering, and expert analysis, but yields no comparable scores. Another line further incorporates scoring learned from human expert ratings. Ratings that are both large-scale and consistent, however, are hard to obtain for PAA, since aesthetic judgment is inherently subjective and each photograph must be rated along multiple photography-specific dimensions beyond an overall score. This raises a central question: *How can we build a PAA model that delivers both grounded analyses and reliable scores with limited high-quality expert ratings?*

To address this question, we propose **AesAlign**. Our key observation is that analyses and ratings require sharply different supervision. Strong multimodal models can generate fine-grained aesthetic analyses at scale, supplying abundant supervision for photographic aesthetic understanding. Reliable ratings, in contrast, require a consistent human scale, and model-generated ratings are often biased. AesAlign therefore learns aesthetic understanding from large-scale analyses, then quantitatively distills it into scores using a small set of human expert ratings. Scores are thus read from the same representations that produce the analyses, grounding them in those analyses while using limited ratings efficiently.

We instantiate this idea with a data pipeline and two-stage training. We first construct AesAlign-500K, comprising 500K smartphone photographs. Following our proposed attribute framework, a multimodal large language model generates fine-grained aesthetic analyses for all 500K images, while experts provide calibrated overall and dimension-level ratings for 50K images. In Stage 1, the model learns to generate fine-grained aesthetic analyses from the MLLM-generated annotations. In Stage 2, we freeze the model and fit lightweight linear regressors on its representations using expert ratings to map the acquired understanding to the human scale. Extensive experiments demonstrate strong scoring performance and cross-dataset generalization.

Our contributions are summarized as follows:

- We formulate PAA as producing fine-grained analyses together with reliable scores, and identify the supervision asymmetry between scalable model-generated analyses and scarce expert ratings as the key to addressing it data-efficiently.
- We propose **AesAlign**, a two-stage framework that learns aesthetic understanding from large-scale analyses and quantitatively distills it into scores through lightweight linear regressors calibrated on a small set of human ratings.
- We construct AesAlign-500K, which provides analyses for 500K smartphone photographs under a proposed attribute framework, along with expert-calibrated for a 50K subset.
- Extensive experiments show that AesAlign achieves strong scoring performance with substantially fewer human ratings than direct score fine-tuning and generalizes across datasets.

2 Related Work

Photography Aesthetics Assessment. Photography Aesthetics Assessment (PAA) aims to evaluate the aesthetic quality of photographs by modeling human aesthetic judgments. As shown in Table 1, early aesthetic benchmarks, such as AVA Murray et al. (2012), primarily focus on overall aesthetic scoring Ren et al. (2017); Hosu et al. (2020); He et al. (2022). AADB Kong et al.

Table 1: Comparison of image aesthetics datasets and related perceptual quality benchmarks. AesAlign combines MLLM-generated analyses for 500K photographs with expert-calibrated overall and attribute-level scores for a 50K subset. Abbreviations are defined below the table.

Dataset	#Images	Attribute Scheme	Text Annotations		Score Annotations	
			Granularity	Source	Granularity	Source
AVA Murray et al. (2012)	255,528	Fixed	–	–	O	Human
AADB Kong et al. (2016)	10,000	Fixed	–	–	O + A	Human
FLICKR-AES Ren et al. (2017)	40,499	None	–	–	O	Human
KonIQ-10K Hosu et al. (2020)	10,073	None	–	–	O	Human
SPAQ Fang et al. (2020)	11,125	Fixed	–	–	O + A	Human
PARA Yang et al. (2022)	31,229	Fixed	–	–	O + A	Human
TAD66K He et al. (2022)	66,000	T	–	–	O	Human
RPCD Vera Nieto et al. (2022)	74,000	Open-ended	Free-form critique	Human	O	MLLM
UNIAA-QA Zhou et al. (2024)	5,354	Fixed	A Question answering	Human	–	–
AesMMIT Huang et al. (2024a)	21,904	Open-ended	O + A Question answering	Human + MLLM	–	–
ArtiMuse-10K Cao et al. (2026)	10,000	Fixed	A Analysis	Human	O	Human
HumanBeauty-50K Liao et al. (2025)	50,022	Hierarchical	–	–	O + A	Human
AesAlign-Base	**450,000**	G + T	O + A Analysis	MLLM	O + A	MLLM
AesAlign-Calibrated & Eval	**50,000**	G + T	O + A Analysis	MLLM	O + A	Human-calibrated

Abbreviations. T = theme-specific; G + T = general + theme-specific; O = overall; A = attribute-level. O + A Analysis and O + A Question answering denote overall and attribute-level analyses and question answering, respectively.

(2016), SPAQ Fang et al. (2020), and PARA Yang et al. (2022) further introduce attribute-level annotations, while RPCD Vera Nieto et al. (2022) and more recent benchmarks incorporate language-based aesthetic analyses Zhou et al. (2024); Huang et al. (2024a); Cao et al. (2026). HumanBeauty-50K Liao et al. (2025) further provides multidimensional numerical supervision. Accordingly, PAA has evolved toward increasingly fine-grained and comprehensive evaluation, extending from holistic score prediction Talebi & Milanfar (2018); He et al. (2022) to multidimensional aesthetic assessment that integrates quantitative scoring with language-based analysis Zhou et al. (2024); Huang et al. (2024b); Cao et al. (2026). Yet, fine-grained quantitative scoring still relies heavily on high-quality human ratings, which are costly to acquire at scale.

Multimodal LLM for scoring. Previous MLLM-based scoring methods generally follow two paradigms: continuous scoring and next-token prediction. Continuous scoring attaches a score head to hidden representations of the LLM Liao et al. (2025); Zhang et al. (2025); Zang et al. (2025); Dai et al. (2024); He et al. (2024); Xu et al. (2024), enabling direct prediction of continuous scores. However, such approaches typically require adapting the learned representations or jointly optimizing additional scoring objectives. Next-token prediction instead represents scores as vocabulary tokens, including digit-token prediction Li et al. (2025); Wang et al. (2025) and level-based scoring Cao et al. (2026); Zhou et al. (2024); Wu et al. (2023; 2024). While naturally compatible with language generation and textual rationales, token-based scoring introduces a discretized output space and requires additional mechanisms to recover fine-grained continuous scores. In our work, we instead adopt lightweight linear probes on frozen aesthetic representations, enabling continuous score prediction without further updating the learned representation.

3 AesAlign-500K Dataset

3.1 Overview

AesAlign-500K comprises 500K smartphone photographs and is randomly partitioned into three disjoint subsets: *AesAlign-Base* (450K), *AesAlign-Calibrated* (45K), and *AesAlign-Eval* (5K), as illustrated in Figure 2. All subsets are annotated with MLLM-generated aesthetic analyses and scores, while the scores in *AesAlign-Calibrated* and *AesAlign-Eval* are further calibrated by experts. *AesAlign-Base* provides large-scale supervision for aesthetic understanding, *AesAlign-Calibrated* supplies expert-calibrated ratings for score calibration, and *AesAlign-Eval* is held out exclusively for quantitative evaluation.

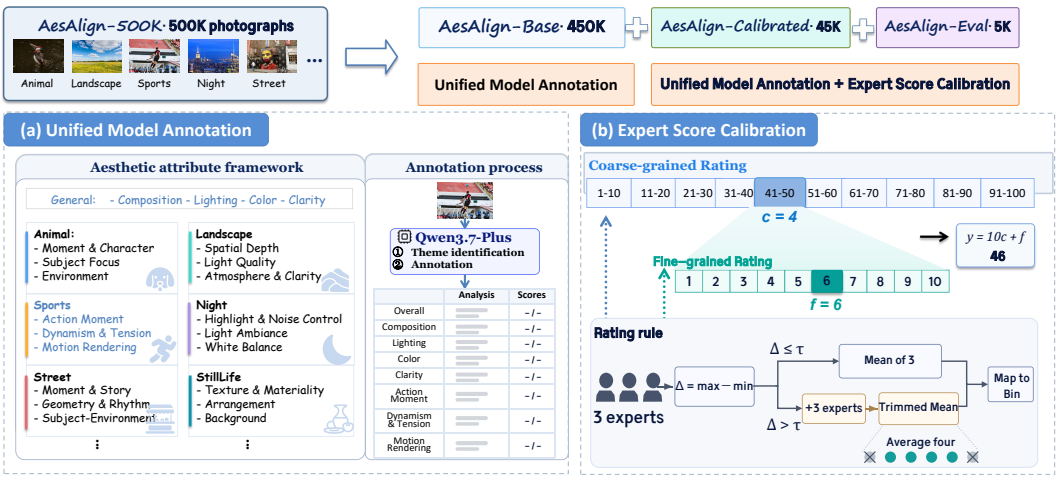


Figure 2: Annotation pipeline of *AesAlign-500K*, comprising *AesAlign-Base* (450K), *AesAlign-Calibrated* (45K), and *AesAlign-Eval* (5K). (a) Following the photography aesthetic attribute framework, Qwen3.7-Plus assigns a photographic theme and generates aesthetic analyses and initial overall and dimension-level scores for all 500K images. (b) For *AesAlign-Calibrated* and *AesAlign-Eval*, experts replace the initial model-generated scores using a coarse-to-fine rating protocol.

3.2 Data Collection

We collect photographs from Flickr captured with 34 smartphone models, covering diverse subjects under unconstrained shooting conditions. We first remove unreadable images and records with missing or invalid metadata. We then filter technically unusable or uninformative samples. Specifically, we exclude images with a longest side below 512 pixels, a grayscale-intensity standard deviation below 3, or no chromatic pixels. To further remove near-duplicates, we extract 512-dimensional SSCD features and retrieve nearest neighbors using FAISS. Pairs with a cosine similarity of at least 0.95 are grouped into near-duplicate clusters, from which we retain only the highest-resolution image. Following filtering and deduplication, 500K photographs remain to form *AesAlign-500K*.

3.3 Photography Aesthetic Attribute Framework

Photography aesthetics depends on both general principles shared across photographs and criteria specific to the subject or shooting scenario. Our framework evaluates every photograph along four general dimensions: Composition, Lighting, Color, and Clarity. It further defines three dedicated dimensions for each of ten photographic themes, with each photograph evaluated on the three corresponding to its theme. For example, Sports complements the general dimensions with Action Moment, Dynamism & Tension, and Motion Rendering, emphasizing how action is captured and conveyed. The complete taxonomy is provided in the appendix.

3.4 Annotation Pipeline

Unified Model Annotation. We use Qwen3.7-Plus to annotate all images in *AesAlign-500K* following the aesthetic attribute framework. For each image, the model first identifies its photographic theme and selects the corresponding three theme-specific dimensions. It then generates eight analysis–score pairs covering overall aesthetics, the four general dimensions, and the three selected theme-specific dimensions. The MLLM-generated scores are retained in *AesAlign-Base* and replaced by expert ratings in *AesAlign-Calibrated* and *AesAlign-Eval*.

Expert Score Calibration. To obtain these expert ratings, we use a *coarse-to-fine hierarchical rating protocol* that decomposes rating on a 1–100 scale into two successive ten-level judgments: coarse-grained quality binning followed by fine-grained rating within the assigned band. The protocol is applied separately to overall aesthetics and each of the seven applicable dimensions. Ex-

perts use shared criterion definitions and human-validated reference anchors, with MLLM-generated scores and analysis texts hidden throughout annotation.

Coarse-grained Rating. We divide the 1–100 scale into ten consecutive ten-point bands, encoded in ascending order as $\mathcal{B} = \{0.5, 1.5, \ldots, 9.5\}$. For each image–criterion item, three experts independently select quality bands, with $b_i \in \mathcal{B}$ denoting the choice of expert i. We compute their disagreement as $\Delta_c = \max_i b_i - \min_i b_i$. When $\Delta_c \leq 2$, we average the three judgments. If $\Delta_c > 2$, three additional experts independently rate the item. We then discard one highest and one lowest value among the six judgments and average the remaining four. The resulting mean $\bar{b}$ determines the coarse-band index $c = \lfloor \bar{b} \rfloor$, corresponding to scores $10c + 1$ through $10c + 10$.

Fine-grained Rating. Within the assigned ten-point interval, three experts independently select fine levels using human-validated anchors for that interval and criterion, with $f_i \in \{1, \ldots, 10\}$ denoting the choice of expert i. We compute their disagreement as $\Delta_f = \max_i f_i - \min_i f_i$. When $\Delta_f \leq 3$, we average the three judgments. If $\Delta_f > 3$, three additional experts independently rate the item within the same band, and we apply the same trimmed aggregation as in the coarse stage. The resulting mean $\bar{f}$ is rounded to the nearest integer to obtain the final fine level, $f = \lfloor \bar{f} + 0.5 \rfloor$.

Final Score Computation. Given the final coarse-band index $c \in \{0, \ldots, 9\}$ and fine level $f \in \{1, \ldots, 10\}$, we compute

$$y = 10c + f. \tag{1}$$

For example, fine level 6 within the 41–50 band yields a score of 46.

4 METHOD

We introduce AesAlign for aesthetic analysis and scoring at both overall and dimension levels. We first formulate the task and supervision (Sec. 4.1), then describe the model architecture (Sec. 4.2) and its two-stage training pipeline (Sec. 4.3).

4.1 PROBLEM FORMULATION

Given a photograph x and an aesthetic evaluation instruction q, Photography Aesthetics Assessment (PAA) aims to produce textual analyses a and numerical scores at both overall and dimension levels. Following the attribute framework in Sec. 3.3, the applicable dimensions are

$$\mathcal{D}(x) = \mathcal{D}_{\text{general}} \cup \mathcal{D}_{\text{theme}}(x), \tag{2}$$

comprising four general dimensions and three dimensions associated with the image's theme. Following the splits in Sec. 3.1, we define $\mathcal{A} = \{(x_i, q_i, a_i)\}_{i=1}^{N_A}$ from AesAlign-Base and $\mathcal{R} = \{(x_i, q_i, \mathbf{y}_i)\}_{i=1}^{N_R}$ from AesAlign-Calibrated, where $\mathbf{y}_i$ contains the overall and applicable dimension-level expert ratings.

4.2 MODEL DESIGN

As illustrated in Figure 3, AesAlign comprises an aesthetic analysis model and a set of independent calibrated scorers. The analysis model is built upon Qwen3.5-2B, with pretrained parameters θ and LoRA parameters ϕ. Given an image x_i and evaluation instruction q_i, the model predicts the photographic theme and generates the corresponding overall and dimension-level aesthetic analyses.

To obtain the representation used for quantitative scoring, we extract the hidden states from a selected LLM layer ℓ and average those corresponding to the instruction-token positions:

$$\mathbf{h}_i = \frac{1}{|\mathcal{T}(q_i)|} \sum_{t \in \mathcal{T}(q_i)} \mathbf{z}_t^{(\ell)}(x_i, q_i; \theta, \phi) \in \mathbb{R}^p, \tag{3}$$

where $\mathcal{T}(q_i)$ denotes the set of instruction-token positions, $\mathbf{z}_t^{(\ell)}$ is the hidden state at position t of layer ℓ, and p denotes the hidden dimension. The resulting $\mathbf{h}_i$ serves as a shared aesthetic representation for score prediction.

We associate an independent calibrated scorer with the overall aesthetic score and each of the 34 predefined aesthetic dimensions. For scoring target k, the corresponding scorer maps $\mathbf{h}_i$ to a scalar

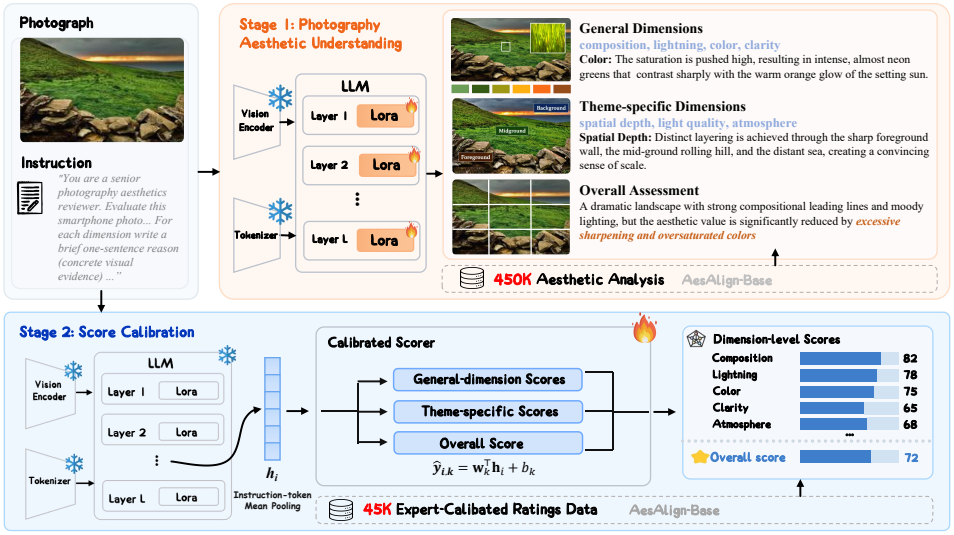


Figure 3: Overview of AesAlign. Stage 1 adapts the multimodal model with LoRA on analyses from *AesAlign-Base* to generate overall and dimension-level aesthetic analyses. Stage 2 freezes the adapted model, extracts its hidden representations, and fits independent scorers using expert ratings from *AesAlign-Calibrated* to produce corresponding scores.

prediction:

$$\hat{y}_{i,k} = \mathbf{w}_k^\top \mathbf{h}_i + b_k, \tag{4}$$

where $\mathbf{w}_k$ and b_k are target-specific parameters. At inference time, the predicted theme selects the three applicable theme-specific scorers, which, together with the overall scorer and four general-dimension scorers, produce eight numerical aesthetic scores from $\mathbf{h}_i$.

4.3 Training Pipeline

The two-stage pipeline exploits the supervision asymmetry between abundant analysis texts and limited expert ratings. Stage 1 uses AesAlign-Base to adapt the analysis model, whereas Stage 2 uses AesAlign-Calibrated to fit the independent calibrated scorers while keeping the analysis model fixed.

Stage 1: Photography Aesthetic Understanding. Using the analysis set $\mathcal{A}$, we keep θ fixed and optimize the LoRA parameters ϕ to generate the target analysis token by token. The training objective is the negative log-likelihood of the analysis sequence:

$$\mathcal{L}_{\text{analysis}}(\phi) = -\frac{1}{N_A} \sum_{i=1}^{N_A} \sum_{t=1}^{T_i} \log p_{\theta,\phi}(a_{i,t} \mid x_i, q_i, a_{i,<t}), \tag{5}$$

where T_i is the length of the target analysis and $a_{i,<t}$ denotes its preceding tokens. This stage yields the adapted parameters $\phi^\star$ and a model capable of generating fine-grained aesthetic analyses.

Stage 2: Score Calibration. We freeze both θ and $\phi^\star$ and extract $\mathbf{h}_i$ using Eq. 3. For each target k, let $\mathcal{I}_k$ denote the indices of samples in $\mathcal{R}$ with an applicable expert rating $y_{i,k} \in [1, 100]$. We fit the corresponding calibrated scorer independently using ridge regression:

$$(\mathbf{w}_k^\star, b_k^\star) = \arg\min_{\mathbf{w}_k, b_k} \sum_{i \in \mathcal{I}_k} \left(y_{i,k} - \mathbf{w}_k^\top \mathbf{h}_i - b_k\right)^2 + \lambda_k \|\mathbf{w}_k\|_2^2, \tag{6}$$

where λ_k is selected independently for target k. Thus, Stage 2 trains only the scorer parameters $(\mathbf{w}_k, b_k)$, leaving the analysis model and its generation capability unchanged.

Table 2: Overall aesthetic scoring on *AesAlign-Eval* and the expert-reannotated AADB and PARA evaluation sets. The evaluated images from AADB and PARA are re-annotated using the same expert rating protocol as *AesAlign-Eval*.

Model	**AADB** Kong et al. (2016)		**PARA** Yang et al. (2022)		**AesAlign-Eval**	
	SRCC ↑	**PLCC** ↑	**SRCC** ↑	**PLCC** ↑	**SRCC** ↑	**PLCC** ↑
Zero-shot						
InternVL3-8B Zhu et al. (2025)	0.684	0.673	0.773	0.859	0.654	0.636
Qwen3.8-Max	0.663	0.711	0.751	0.825	0.668	0.698
Gemini-3 flash	0.635	0.656	0.715	0.760	0.612	0.646
MUSIQ Ke et al. (2021)	0.451	0.494	0.599	0.720	0.189	0.242
Q-Align Wu et al. (2023)	0.656	0.675	0.708	0.720	0.581	0.610
ArtiMuse Cao et al. (2026)	0.668	0.671	0.724	0.748	0.648	0.675
Supervised						
InternVL3-8B Zhu et al. (2025)	0.670	0.708	0.748	0.819	0.662	0.691
MUSIQ Ke et al. (2021)	0.643	0.669	0.691	0.751	0.563	0.596
Q-Align Wu et al. (2023)	0.675	0.673	0.748	0.704	0.601	0.628
AesAlign (Ours)	**0.732**	**0.752**	**0.816**	**0.875**	**0.779**	**0.800**

5 Experiments

5.1 Implementation Details

We adopt pretrained Qwen3.5-2B as our base model. During text pretraining on *AesAlign-Base* (450K), we use a batch size of 60 (per-device batch size 2 with gradient accumulation 15 on 2 GPUs) and a learning rate of 4e-5 for 1 epoch, with LoRA (rank 128, $\alpha = 256$, dropout 0.05) applied to all linear layers and the vision–language merger, keeping the vision tower frozen. In Stage 2, we freeze the model and fit a Ridge probe on the expert-calibrated scores of *AesAlign-Calibrated* (45K), using the mean-pooled layer-22 hidden states of the text tokens from a single forward pass, with per-dimension standardization, and selecting the regularization strength through leave-one-out cross-validation on the fitting set over 13 log-spaced values of $\alpha \in [10^{-1}, 10^{5}]$ (selected per target). Stage 1 uses AdamW with a cosine learning rate schedule (warmup ratio 0.05, bf16 mixed precision). Training is conducted on 2 NVIDIA RTX 5090D (24GB) GPUs, taking approximately 14.5 and 1 hours for the two stages, respectively.

5.2 Aesthetic Scoring Performance

We evaluate overall aesthetic scoring on *AesAlign-Eval* and two external photographic datasets, AADB and PARA. Since AADB and PARA use different rating protocols and scales, we re-annotate their test sets using the expert calibration protocol in Sec. 3.4 for consistent cross-dataset evaluation. We compare three categories of methods: general-purpose MLLMs (InternVL3-8B Zhu et al. (2025), Qwen3.8-Max, and Gemini-3 Flash), traditional IAA models (MUSIQ Ke et al. (2021)), and aesthetics-specialized MLLMs (Q-Align Wu et al. (2023) and ArtiMuse Cao et al. (2026)). We report SRCC and PLCC under zero-shot evaluation and supervised rating learning on *AesAlign-Calibrated*, following each method's training procedure.

As shown in Table 2, general-purpose MLLMs exhibit strong zero-shot aesthetic knowledge, whereas specialized IAA models transfer less effectively to PAA due to differences in scoring criteria and image sources. Supervised learning improves performance on *AesAlign-Eval*, but these improvements do not translate into clear benefits on the external datasets; InternVL3-8B even performs worse after adaptation. AesAlign performs best on all three datasets, reaching 0.779 SRCC and 0.800 PLCC on *AesAlign-Eval* and demonstrating strong cross-dataset generalization.

5.3 Sample Efficiency of Score Adaptation

To evaluate the sample efficiency of score calibration, we compare two approaches after Stage 1: Score SFT updates the model through LoRA to generate numerical scores, while the calibrated scorer fits Ridge regression on frozen representations. Both methods use identical nested subsets of *AesAlign-Calibrated* and are evaluated on *AesAlign-Eval*.

Table 3: Sample efficiency of score calibration. Both methods use the same Stage 1 checkpoint.

Method	Metric	Expert-Calibrated Images						
		450 1%	**900** 2%	**2,250** 5%	**4,500** 10%	**9,000** 20%	**22,500** 50%	**45,000** 100%
Score SFT (LoRA)	SRCC ↑	0.686	0.691	0.694	0.694	0.697	0.699	0.706
	PLCC ↑	0.726	0.730	0.731	0.733	0.737	0.738	0.741
Calibrated Scorer (Ridge)	SRCC ↑	0.739	0.741	0.751	0.753	0.759	0.768	0.779
	PLCC ↑	0.769	0.771	0.778	0.776	0.779	0.783	0.800

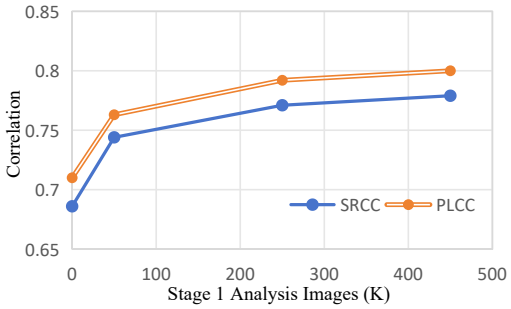


Figure 4: Effect of Stage 1 data scale.

Table 4: Model structure ablation on *AesAlign-Eval*. Default settings are shown in bold.

Component	Setting	SRCC ↑	PLCC ↑
Backbone	Qwen3.5-0.8B	0.740	0.768
	Qwen3.5-2B	0.779	0.800
	Qwen3.5-4B	0.781	0.802
Score Head	Linear	0.729	0.747
	Ridge	0.779	0.800
	2-Layer MLP	0.764	0.780
Probe Feature	Last Token	0.774	0.787
	Mean Pooling	0.779	0.800

Table 5: Ablation of two-stage learning.

Stage 1	Score Calibration	SRCC ↑	PLCC ↑
✗	Score SFT (LoRA)	0.684	0.705
✗	Calibrated Scorer (Ridge)	0.686	0.710
✓	Score SFT (LoRA)	0.706	0.741
✓	**Calibrated Scorer (Ridge)**	**0.779**	**0.800**

Table 6: Ablation of attribute framework.

Stage 1 Supervision	SRCC ↑	PLCC ↑
None	0.686	0.710
Overall	0.726	0.748
Overall + General	0.757	0.779
Overall + General + Theme-Specific	**0.779**	**0.800**

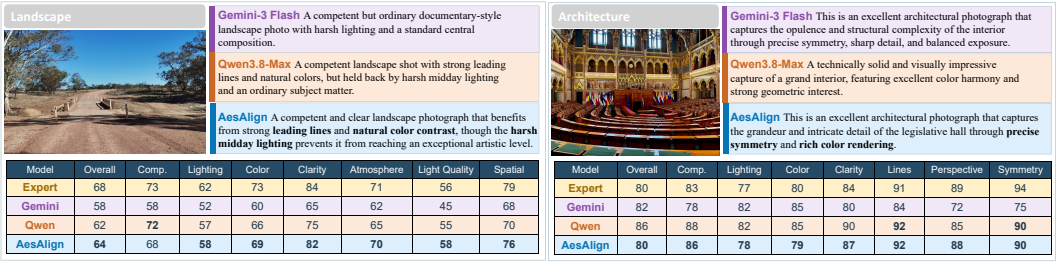
As shown in Table 3, the calibrated scorer consistently outperforms Score SFT across all data scales. Notably, with only 450 expert-calibrated images, it surpasses Score SFT trained on all 45,000 images. Moreover, Score SFT exhibits only limited gains as the amount of calibration data increases, indicating substantially lower sample efficiency for score calibration. These results support separating aesthetic representation learning from score calibration, with a lightweight readout mapping frozen features to expert ratings.

5.4 Ablation Studies

Scaling Analysis Supervision To examine the effect of analysis data scale, we vary the number of Stage 1 training images from 0 to 450K and evaluate the resulting scoring performance on *AesAlign-Eval*. Figure 4 shows that scoring performance generally improves as the amount of aesthetic analysis supervision increases. This suggests that larger-scale analysis training helps learn representations that better support aesthetic scoring.

Model Design As shown in Table 4, Qwen3.5-2B improves over 0.8B, whereas 4B brings only a marginal further gain, suggesting diminishing returns from increasing backbone size. Ridge outperforms both unregularized Linear and the 2-Layer MLP. The large gap with Linear is consistent with regularization being important for the frozen features, while the MLP result suggests that a more complex mapping is unnecessary here. Mean pooling slightly outperforms the last-token feature, suggesting that aggregating information from all instruction-token positions is more effective for scoring than relying on the final token alone.

Training Stage Ablation To examine the contributions of the two stages, we compare Score SFT and the calibrated scorer with and without Stage 1 training, using the same backbone and expert ratings. As shown in Table 5, the two methods perform similarly without Stage 1. Adding Stage 1 improves both methods, with SRCC/PLCC increasing from 0.684/0.705 to 0.706/0.741 for Score



Model	Overall	Comp.	Lighting	Color	Clarity	Atmosphere	Light Quality	Spatial
Expert	68	73	62	73	84	71	56	79
Gemini	58	58	52	60	65	62	45	68
Qwen	62	**72**	57	66	75	65	55	70
AesAlign	**64**	68	**58**	**69**	**82**	**70**	**58**	**76**

Model	Overall	Comp.	Lighting	Color	Clarity	Lines	Perspective	Symmetry
Expert	80	83	77	80	84	91	89	94
Gemini	82	78	82	85	80	84	72	75
Qwen	86	88	82	85	90	**92**	85	**90**
AesAlign	**80**	**86**	**78**	**79**	**87**	**92**	**88**	**90**

Figure 5: Examples of aesthetic analyses and scores compared with expert ratings.



Figure 6: Examples of consistency between aesthetic analyses and scores.

SFT and from 0.686/0.710 to 0.779/0.800 for the calibrated scorer. The larger gains with the calibrated scorer indicate that its frozen-representation readout makes more effective use of the aesthetic representations learned in Stage 1.

Attribute Framework To isolate the contribution of attribute-level analysis supervision, we progressively add general and theme-specific analyses to Stage 1 while keeping score calibration fixed. As shown in Table 6, overall analysis training improves upon the pretrained representation, and adding general and theme-specific analyses brings further gains. The full setting reaches 0.779 SRCC and 0.800 PLCC, suggesting that general and theme-specific analyses both help Stage 1 learn representations useful for overall aesthetic scoring.

5.5 Qualitative Comparison of Aesthetic Analyses and Scores

Comparison with Expert Ratings Figure 5 shows two photographs with their expert ratings and model predictions. In both examples, AesAlign's overall and dimension-level scores are closer to the expert ratings than those of the compared models. The accompanying analyses provide context for interpreting the predictions in terms of visible attributes, including composition, lighting, and theme-specific criteria.

Consistency between Aesthetic Analyses and Scores To compare how consistently different models align their aesthetic analyses with their scores, Figure 6 presents two examples: one for overall aesthetics and one for the color dimension. Across both examples, all models provide positive assessments and analyses. AesAlign's relatively high scores closely match the positive judgments in its analyses, illustrating consistency between its textual assessments and score predictions.

6 Conclusion

In this paper, we introduce AesAlign-500K, a dataset of 500K smartphone photographs with fine-grained aesthetic analyses covering overall, general, and theme-specific criteria. A 50K-image subset further receives expert-calibrated ratings under a consistent scoring protocol. Building on this dataset, we propose AesAlign, a two-stage framework that learns photographic aesthetic understanding from large-scale analyses and then maps frozen representations to expert ratings using lightweight regressors. AesAlign achieves strong scoring performance with substantially fewer human ratings than direct score fine-tuning, generalizes across datasets, and preserves fine-grained analysis capabilities.

GENERATIVE AI DISCLOSURE

In this work, we used generative AI tools to generate synthetic datasets, assist with the design and refinement of research methodology and experiments, support translation, and clean and reformat datasets. We did not use generative AI tools for other tasks requiring mandatory disclosure, or such tasks were not applicable to this work.

Additionally, we used generative AI tools to summarize and analyze existing literature and to suggest the structure of the paper. All AI-assisted outputs were reviewed and verified by the authors prior to inclusion in the final manuscript. The authors take full responsibility for the final content of this work, including all text, claims, data, and artifacts produced with the aid of generative AI.

REFERENCES

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. 3, 7

Juntao Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *Advances in Neural Information Processing Systems*, 37:17161–17214, 2024. 3

Niladri Shekhar Dutt, Duygu Ceylan, and Niloy J Mitra. Monetgpt: Solving puzzles enhances mllms' image retouching skills. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025. 2

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3677–3686, 2020. doi: 10.1109/CVPR42600.2020.00373. 3

Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 309–316, 2016. 1, 2

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pp. 942–948, 2022. 2, 3

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024. 3

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. 2, 3

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920, 2024a. 3

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*, 2024b. 3

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021. 7

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. 2, 3, 7

Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025. 3

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. 3

Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. *arXiv preprint arXiv:2509.21871*, 2025a. 1

Henglin Liu, Nisha Huang, Chang Liu, Jiangpeng Yan, Huijuan Huang, Jixuan Ying, Tong-Yee Lee, Pengfei Wan, and Xiangyang Ji. Bridging cognitive gap: Hierarchical description learning for artistic image aesthetics assessment. *arXiv preprint arXiv:2512.23413*, 2025b. 1

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. 2, 3

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer's eye: Teaching multimodal large language models to see, and critique like photographers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24807–24816, 2025. 2

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 638–647, 2017. doi: 10.1109/ICCV.2017.77. 2, 3

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 3

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. 3

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 3

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5437–5446, 2018. 2

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*, 2023. 3, 7

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, volume 2024, pp. 12547–12573, 2024. 3

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 3

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19861–19869, 2022. 3, 7

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025. 3

Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Haotian Wang, Kai Wu, Bo Cui, et al. Basereward: A strong baseline for multimodal reward model. *arXiv preprint arXiv:2509.16127*, 2025. 3

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. 3

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. 7

A Appendix

A.1 AesAlign-500K Dataset Details

A.1.1 Photography Aesthetic Attributes

Tables 7 and 8 summarize the four general dimensions and the three dimensions defined for each photographic theme.

Table 7: General aesthetic dimensions evaluated for every photograph.

Dimension	**Description**
Composition	Subject placement, visual balance, negative space, and visual guidance.
Lighting	Exposure and the use of light and shadow to create depth and model the subject.
Color	Color harmony, white balance, and visual appeal.
Clarity	Subject sharpness, noise control, and overall crispness.

Table 8: Theme-specific aesthetic dimensions.

Theme	**Dimension**	**Description**
Portrait	Expression & Mood	Facial expression when visible; otherwise mood conveyed by posture, silhouette, or atmosphere.
	Subject-Background Relation	How the background separates, supports, or provides context for the subject.
	Pose & Framing	The subject's pose and placement within the frame.
Landscape	Spatial Depth	Organization of foreground, midground, and background, and the resulting sense of depth.
	Light Quality	Effective use of dramatic light or well-handled ordinary and overcast light.
	Atmosphere & Clarity	The scene's mood and the visibility of its landscape features.
Street	Moment & Story	A decisive moment or the narrative atmosphere of the place.
	Geometry & Rhythm	Visual structure formed by lines, shapes, and repetition.
	Subject-Environment	The relationship between subjects and their surroundings.
Sports	Action Moment	Timing and capture of a meaningful point in the action.
	Dynamism & Tension	A sense of motion, anticipation, or static tension.
	Motion Rendering	Effective depiction of motion through a sharp freeze, intentional blur, or panning.
Night	Highlight & Noise Control	Preservation of highlights and control of visible noise.
	Light Ambiance	The mood created by available and artificial light.
	White Balance	Handling of color casts and mixed lighting.

Theme	Dimension	Description
Food	Appetite Appeal	How inviting and appetizing the food appears.
	Presentation	Deliberate plating or an appealing natural arrangement.
	Lighting Quality	How lighting reveals the food's shape, color, and texture.
Macro	Subject Sharpness	Sharpness and visible detail in the key subject region.
	Subject Isolation	Separation through background blur or a clean, fill-frame composition.
	Texture	Visibility and rendering of fine surface detail.
Architecture	Lines & Geometry	Use of structural lines and geometric forms.
	Perspective Control	Handling of perspective, vertical lines, and spatial proportions.
	Symmetry & Balance	Symmetric organization or visual balance of architectural elements.
StillLife	Texture & Materiality	Rendering of surface texture and material properties, such as glass, metal, wood, or fabric.
	Arrangement	Placement of objects and their visual relationships.
	Background	How the backdrop supports the objects without distraction.
Animal	Moment & Character	An expressive moment or character conveyed through posture or silhouette.
	Subject Focus	Eye sharpness when eyes are visible; otherwise sharpness of the key body part.
	Environment	How the animal is presented within its surroundings.

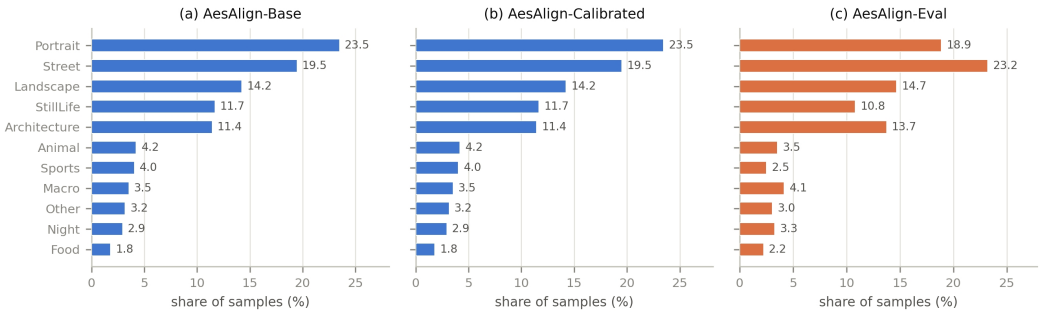


Figure 7: Theme distributions of *AesAlign-Base*, *AesAlign-Calibrated*, and *AesAlign-Eval*. Values indicate the percentage of images in each subset.

A.1.2 Dataset Statistics

Figure 7 shows the distribution of the ten photographic themes and the Other category across *AesAlign-Base*, *AesAlign-Calibrated*, and *AesAlign-Eval*. Portrait and Street have the largest shares, while every category is represented in the evaluation set.

Figure 8 compares the initial LLM scores with expert-calibrated scores on *AesAlign-Eval*. Relative to the LLM scores, the calibrated scores place more images in the 50–60 range and fewer in the higher score ranges, illustrating the effect of expert calibration on the rating distribution.

A.1.3 Annotation Prompt Templates

We use a single prompt to obtain the photographic theme, aesthetic analyses, and initial scores for each image. For readability, the complete prompt is divided below into consecutive parts. The model receives these parts as one instruction and returns a JSON object. Expert calibration subsequently replaces the initial scores in *AesAlign-Calibrated* and *AesAlign-Eval*.

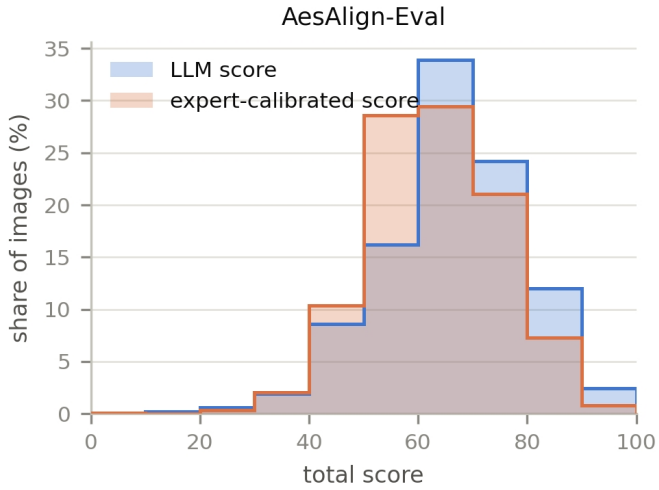


Figure 8: Distributions of initial LLM-generated scores and expert-calibrated scoring labels on *AesAlign-Eval*.

Annotation prompt: image screening and theme selection

```
You are a senior photography aesthetics reviewer. Evaluate this
    smartphone photo. Follow the steps and output ONLY a JSON object
    .

STEP 1---reject_flag (choose one; marks images to FILTER OUT before
     use):
- "none": normally usable
- "harmful": pornographic, gore/violence, hateful, or otherwise
    harmful content
- "watermark": has ANY watermark, stock-photo logo, signature, date
    /time stamp, copyright or URL text, or semi-transparent overlaid
     text/logo---INCLUDING small, faint, or corner/edge-placed ones
    (look carefully at corners and edges). Do NOT count natural in-
    scene text such as shop signs, billboards, or product labels. Do
     NOT count phone-native/OEM camera-signature watermarks
    automatically added by the capture device---e.g. "Shot on Huawei
     / Leica" dual-line camera stamps, "Samsung Galaxy" signatures,
    or similar built-in device/camera-model captions; these mark the
     shooting device and should be treated as "none".
(If both harmful and watermark apply, choose "harmful".) Still
    complete the scoring regardless of reject_flag.

STEP 2---theme (choose EXACTLY one). Apply these disambiguation
    rules carefully:
- Portrait: a person/people are the CLEAR MAIN subject; focus is on
     face/expression/pose/styling; background is secondary. (A
    single person filling much of the frame → Portrait.)
- Street: public/urban or everyday-life scenes; if people appear
    they are ELEMENTS of the scene (candid, environmental), NOT the
    sole subject; emphasis is on the moment/atmosphere/geometry of
    the place. (A person walking through a street scene → Street,
    not Portrait.)
- Sports: athletic or physical-action moments (running, ball games,
     cycling, skating, extreme sports, dance); emphasis on motion,
    action, decisive timing. (A person in athletic motion → Sports,
    not Portrait.)
```

```
- Landscape: natural open scenery (mountains, sea, fields, sky)
    with no dominant human subject.
- Night: scenes defined by night / low-light / artificial lighting.
- Food: food or drink as the main subject.
- Macro: extreme close-up of small subjects (insects, flowers,
    textures, fine detail).
- Architecture: buildings/structures as main subject; emphasis on
    form and lines.
- StillLife: deliberately arranged inanimate objects.
- Animal: animal(s) as the main subject.
- Other: none of the above.
```

Annotation prompt: overall and general-dimension scoring

```
STEP 3---scoring. ALL scores are INTEGERS 0-100.

ABSOLUTE TOTAL SCORE anchors (total_score, integer 0-100, ONE
    unified scale across ALL themes):
- 90-100 Exceptional: composition/light/color/subject in high
    synergy; strong impact or unique vision; professional level
- 75-89 Excellent: clearly intentional work; careful composition;
    solid technique; memorable
- 60-74 Good: competent photo; no obvious flaws but ordinary/plain
- 40-59 Fair: visible flaws (loose composition / flat light / weak
    subject)
- 20-39 Poor: multiple clear problems (clutter / exposure errors /
    unclear subject)
- 1-19 Failed: technically failed or no aesthetic value

CRITICAL: the same score means the same overall aesthetic level in
    ANY theme. Do NOT rank within a theme; use one absolute scale.

GENERAL dimensions---score ALL FOUR (integer 0-100):
- Composition: subject placement, balance, negative space, visual
    guidance
- Lighting: exposure correctness, light/shadow layering and
    modeling
- Color: harmony, white balance, appeal
- Clarity: subject sharpness, noise, overall crispness
```

Annotation prompt: theme-specific dimensions

```
THEME-SPECIFIC dimensions---output EXACTLY the THREE dimensions for
     the chosen theme, using these EXACT keys. Do NOT use the theme
    name as a key. Do NOT output fewer than three. For "Other",
    output {}.

- Portrait: "Expression & Mood" (facial expression if a face is
    visible; otherwise mood conveyed by posture/silhouette/
    atmosphere), "Subject-Background Relation", "Pose & Framing"
- Landscape: "Spatial Depth", "Light Quality" (golden-hour/dramatic
     light, OR well-handled ordinary/overcast light), "Atmosphere &
    Clarity"
- Street: "Moment & Story" (a decisive moment OR the narrative
    atmosphere of the place), "Geometry & Rhythm", "Subject-
    Environment"
- Sports: "Action Moment", "Dynamism & Tension" (sense of motion,
    OR static tension/anticipation), "Motion Rendering" (whether
```

```
    motion is well conveyed---a crisp freeze of the peak action OR
    intentional motion-blur/panning that conveys speed)
- Night: "Highlight & Noise Control", "Light Ambiance", "White
    Balance"
- Food: "Appetite Appeal", "Presentation" (deliberate plating OR
    appealing natural presentation), "Lighting Quality"
- Macro: "Subject Sharpness", "Subject Isolation" (background
    separation/blur OR clean fill-frame focus), "Texture"
- Architecture: "Lines & Geometry", "Perspective Control", "
    Symmetry & Balance"
- StillLife: "Texture & Materiality" (rendering of surface texture
    and material realism---glass clarity, metal sheen, wood/fabric
    grain), "Arrangement", "Background"
- Animal: "Moment & Character" (expressive moment OR character from
     posture/silhouette), "Subject Focus" (eye sharpness if eyes
    visible; otherwise sharpness of the key body part), "Environment
    "
```

Annotation prompt: rarity, reasoning, and output format

```
STEP 4---rarity (integer 0--3, INDEPENDENT signal for downstream
    filtering; it MUST NOT change total_score or any dimension score
    ): how much the photo shows exceptional creativity, difficulty,
    or rare visual conditions---e.g. extreme weather, aurora/milky-
    way/rainbow/lightning, a once-in-a-moment decisive timing, an
    ingenious/surreal composition, a rare technique. 0 = ordinary (
    DEFAULT for the vast majority of photos); 1 = mildly notable; 2
    = clearly rare/creative; 3 = exceptional/once-in-a-lifetime. Do
    NOT inflate---most photos are 0.

RULES:
- For each dimension write a brief one-sentence reason (concrete
    visual evidence) THEN an integer score 0--100.
- Score each dimension independently; do not give all the same
    value; do not default to high scores.
- total_score follows the absolute anchors; it is an independent
    judgment, NOT a template average of dimensions.
- "thematic" MUST contain exactly the three keys for the chosen
    theme (or {} for Other).
- Write "total_reason" (ONE sentence summarizing the overall
    aesthetic judgment) BEFORE total_score---reason first, then
    score.
- "rarity" is an INDEPENDENT signal and MUST NOT influence
    total_score or the dimension scores in any way.
- Output ONLY valid JSON. No markdown, no extra text.

OUTPUT FORMAT: { "theme": "<Portrait|Landscape|Street|Sports|Night|
    Food|Macro|Architecture|StillLife|Animal|Other>", "reject_flag":
     "<none|harmful|watermark>", "general": { "Composition": {"
    reason": "..", "score": 0}, "Lighting": {"reason": "..", "score
    ": 0}, "Color": {"reason": "..", "score": 0}, "Clarity": {"
    reason": "..", "score": 0} }, "thematic": { "<key1>": {"reason":
     "..", "score": 0}, "<key2>": {"reason": "..", "score": 0}, "<
    key3>": {"reason": "..", "score": 0} }, "total_reason": "one-
    sentence overall aesthetic judgment", "total_score": 0, "rarity
    ": {"reason": "one-sentence why", "score": 0} }
```

Table 9: Correlation between original benchmark ratings and expert re-annotations.

Dataset	PLCC ↑	SRCC ↑
AADB	0.6945	0.6914
PARA	0.8618	0.7834

A.2 Details of Expert Ratings

Assessment System Development. Before large-scale annotation, a panel of photography experts jointly established the assessment system for overall aesthetics and dimension-level ratings. It comprises four general dimensions applicable across themes—Composition, Lighting, Color, and Clarity—and three theme-specific dimensions for each of the ten photographic themes. The criterion definitions are provided in Tables 7 and 8.

Expert Panel and Annotation Procedure. We recruited 50 experts with at least three years of professional experience in visual content production or post-production. Their backgrounds included video editing, film and image color grading, photography, and related fields. Different subsets of the experts completed the coarse- and fine-grained rating stages. Fine-stage raters were not shown the item's assigned coarse band. All raters were blinded to the MLLM-generated analyses and scores.

For each image–criterion item, three experts independently assigned a coarse score band. Items with a range greater than two bands were escalated to three additional experts. Within the resulting band, three experts independently assigned a fine score from 1 to 10; items with a fine-level range greater than three were likewise escalated to three additional experts. For escalated items, we discarded the highest and lowest judgments and averaged the remaining four; otherwise, we averaged the three judgments. The final 1–100 score was computed as described in Sec. 3.4.

Reference Anchor Construction and Validation. We selected candidate images from the photographs to be annotated. For each criterion, multiple experts independently rated the candidates using the shared rubric. We retained candidates whose coarse-band ratings differed by no more than one band and selected five representative anchor images for each coarse score band. Within each coarse band, experts then rated candidates at the fine-grained level. We retained candidates whose fine-level ratings differed by no more than one level and selected three representative anchor images for each fine score level. Three experts independently reviewed the selected anchors for rating appropriateness and representativeness. After confirmation, the anchors were used to guide annotation of the remaining images.

A.3 Additional Implementation Details

Agreement between Original and Re-annotated Ratings To assess how expert re-annotations relate to the original benchmark ratings, we compute PLCC and SRCC between the original scores and our re-annotated scores on the AADB and PARA test images used in our evaluation. Table 9 reports the correlations. The re-annotated scores are positively correlated with the original ratings on both datasets.

Probe Readout Selection We define a probe readout by its pooling method and decoder layer. Features are extracted from a single frozen VLM forward pass and pooled over the designated instruction-token positions. To select the layer, we fit a multi-output Ridge probe at each candidate layer from 2 to 24. The regularization strength is chosen by cross-validation on the training split, and the layer with the highest SRCC on a held-out validation split is selected. Both choices use only *AesAlign-Calibrated*; *AesAlign-Eval* is not used for selection.

A.4 Additional Experimental Results

A.4.1 Dimension-Level Scoring

Tables 10 and 11 report dimension-level scoring on *AesAlign-Eval*.

Table 10: Scoring results for the four general dimensions on *AesAlign-Eval*.

Dimension	**SRCC ↑**	**PLCC ↑**	**MAE ↓**
Composition	0.886	0.887	3.23
Lighting	0.883	0.890	3.31
Color	0.891	0.893	3.29
Clarity	0.832	0.883	3.27

Table 11: Theme-specific scoring results on *AesAlign-Eval*.

Dimension	**SRCC ↑**	**PLCC ↑**	**MAE ↓**
Portrait			
Expression & Mood	0.849	0.843	3.44
Subject-Background Relation	0.835	0.836	3.34
Pose & Framing	0.782	0.789	3.83
Street			
Moment & Story	0.806	0.816	4.26
Geometry & Rhythm	0.911	0.913	3.15
Subject-Environment	0.884	0.888	3.36
Landscape			
Spatial Depth	0.828	0.833	3.58
Light Quality	0.927	0.929	3.42
Atmosphere & Clarity	0.898	0.901	3.09
Sports			
Action Moment	0.716	0.702	4.69
Dynamism & Tension	0.814	0.813	4.20
Motion Rendering	0.570	0.699	3.90
Night			
Highlight & Noise Control	0.927	0.945	4.03
Light Ambiance	0.884	0.904	4.47
White Balance	0.891	0.901	4.32
Food			
Appetite Appeal	0.760	0.825	5.64
Presentation	0.784	0.840	4.54
Lighting Quality	0.776	0.798	4.21
Architecture			
Lines & Geometry	0.845	0.868	3.63
Perspective Control	0.747	0.796	5.05
Symmetry & Balance	0.861	0.896	4.01
StillLife			
Texture & Materiality	0.783	0.785	4.42
Arrangement	0.857	0.856	4.10
Background	0.855	0.869	3.79
Animal			
Moment & Character	0.833	0.849	4.45
Subject Focus	0.845	0.896	4.53
Environment	0.847	0.845	4.00
Macro			
Subject Sharpness	0.818	0.853	4.23
Subject Isolation	0.832	0.837	4.10
Texture	0.842	0.863	4.33

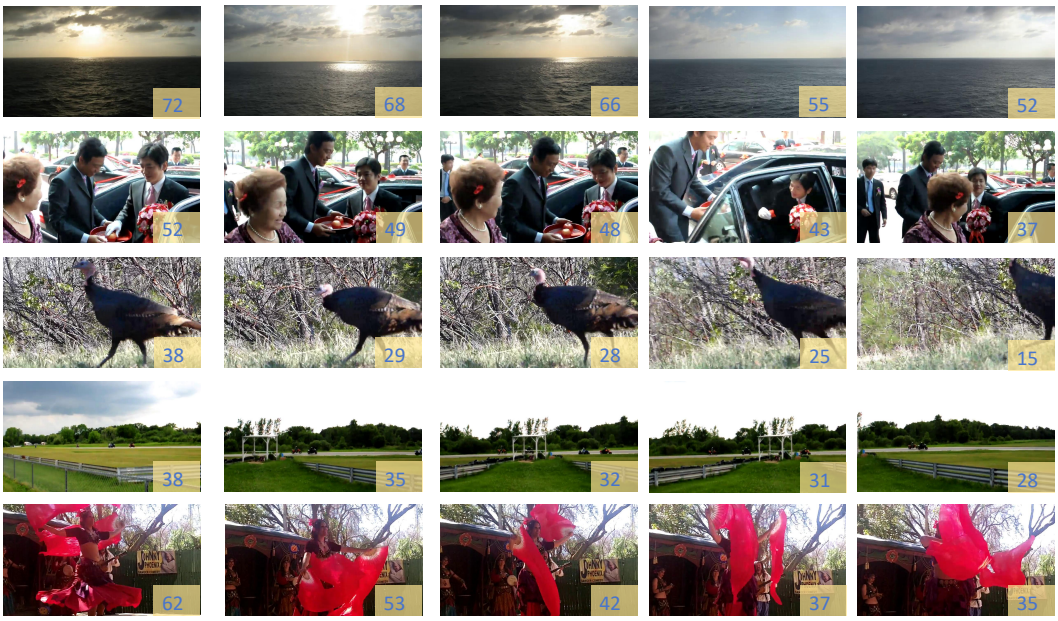


Figure 9: Video frame selection with AesAlign. Sampled frames are ranked by their predicted overall aesthetic scores; the selected frame is highlighted in each sequence.

A.4.2 Video Frame Selection with AesAlign

We apply AesAlign to sampled video frames and rank them by their predicted overall aesthetic scores. Figure 9 shows example frames alongside their scores, with the highest-scoring frame selected from each sequence. The examples illustrate how AesAlign can be used to identify aesthetically appealing moments in video.