AesMind: Harnessing Multimodal Large Language Models for Neuroaesthetic Understanding and Assessment

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in multimodal large language models (MLLMs) have significantly enhanced the interpretability and generalization of image aesthetic assessment (IAA). However, existing methods still predominantly evaluate aesthetics through either a single holistic score or a generic critique, failing to provide fine-grained attribute ratings and structured multidimensional reasoning. To address these limitations, we draw upon *neuroaesthetics*, which explores the cognitive and neural mechanisms of aesthetic appreciation, to establish a comprehensive aesthetic experience taxonomy encompassing the *sensory-motor*, *emotion-valuation*, and *knowledge-meaning* systems. Based on this theoretical framework, we first introduce **NeuroAes**, a large-scale dataset specifically tailored for computational photographic NeuroAesthetics, which comprises 25K images, annotated with over *1.2M fine-grained neuroaesthetic labels* across 48 subcategories and over *2.25M human ratings* spanning the valence-arousal-dominance (VAD) dimensions for both aesthetic and evoked emotions. In particular, NeuroAes is partitioned into a 5K-image **NeuroAes-Bench** for comprehensive aesthetic evaluation and a 20K-image **NeuroAes-Data** for large-scale training. Building upon this foundation, we develop **AesMind**, an all-in-one cognitive aesthetic MLLM that unifies perception, assessment, and description capabilities through a progressive three-stage training paradigm of skill harnessing, supervised fine-tuning (SFT), and group relative policy optimization (GRPO). Extensive experimental results demonstrate that AesMind achieves state-of-the-art performance across all benchmark tasks. Notably, it exhibits superior continuous-emotion regression, unprecedented cross-dataset generalization, and the unique capability to give aesthetic improvement suggestions, proving its ability to deeply understand photographic aesthetics rather than merely reciting superficial visual cues. Both NeuroAes and AesMind will be released to facilitate further research.

1 Introduction

Understanding photographic aesthetics is increasingly crucial for applications ranging from intelligent album management to human-centered image generation. In these scenarios, computational models need to not only judge visual appeal but also identify underlying photographic factors, characterize evoked emotions, and provide actionable feedbacks. Historically, image aesthetic assessment (IAA) has been mainly framed as a data-driven score regression or classification task, predicting mean opinion scores (MOSs) or discrete aesthetic categories (Murray et al., 2012; He et al., 2022). While some works have incorporated specific photographic attributes (Kong et al., 2016; Huang et al., 2024a), these conventional methods inherently compress complex aesthetic evaluation into an isolated scalar score, struggling to model multidimensional human preferences or provide structured reasoning.

Recently, multimodal large language models (MLLMs) have extended IAA from numeric regression toward visual question answering and critique generation (Huang et al., 2024b; Zhou et al., 2024; Cao et al., 2025). However, current MLLM-based IAA methods suffer from a critical bottleneck, *i.e.*, their evaluation dimensions rely primarily on heuristic designs and lack a unified theoretical foundation. Existing datasets are often coarsely annotated and treat aesthetic preference and visual emotion as isolated problems. Consequently, the developed models generally only produce fluent but generic critiques, failing to engage deeply with fine-grained photographic perception and complex emotional valuation.

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.

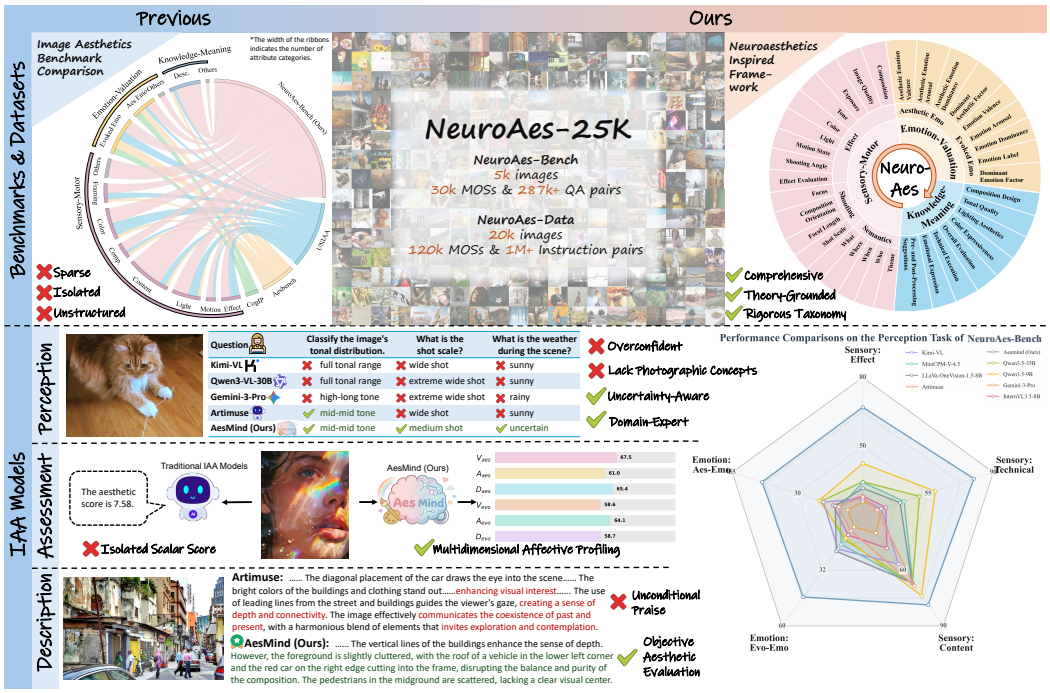


Figure 1: Overview of the proposed computational photographic aesthetics framework. Unlike previous heuristic benchmarks, NeuroAes explicitly unifies sensory perception, affective assessment, and professional description critiques based on neuroaesthetic principles including sensory-motor, emotion-valuation, and knowledge-meaning. Powered by this framework, our AesMind model demonstrates superior multi-task capabilities compared to existing state-of-the-art multimodal models.

To address these limitations, we establish a comprehensive photographic aesthetics taxonomy based on neuroaesthetics. Neuroaesthetics investigates the cognitive and neural mechanisms of aesthetic appreciation and conceptualizes the human aesthetic experience through three interacting cognitive systems including the sensory-motor, emotion-valuation, and knowledge-meaning systems. (Chatterjee & Vartanian, 2014). Within the emotion-valuation system, psychological research further distinguishes aesthetic emotions (induced by photographic execution) from general evoked emotions (elicited by depicted content) (Menninghaus et al., 2019). Therefore, in this paper, based on neuroaesthetic principle, we further formulate the three systems of neuroaesthetics for photographic aesthetics using fine-grained attributes, aesthetic and evoked emotions, and structured critiques. This establishes a unified cognitive chain from perceiving visual attributes, valuing their affective impact to explaining the aesthetic judgment.

Based on this framework, we construct a 25K-image photographic aesthetics dataset, termed **NeuroAes**. To comprehensively understand and evaluate neuroaesthetics, we collect more than *1.2M fine-grained neuroaesthetic labels* covering 48 subcategories and over *2.25M human ratings* on the valence-arousal-dominance (VAD) space (Warriner et al., 2013) for both aesthetic and evoked emotions. The dataset is divided into a 5K-image evaluation benchmark, named **NeuroAes-Bench** and a 20K image training set, termed **NeuroAes-Data**. The NeuroAes-Bench evaluates the unified cognitive chain across three task formats, including the perception of sensory attributes, continuous assessment of VAD scores, and description of aesthetic reasoning, while the NeuroAes-Data provides over one million QA pairs and mean opinion scores (MOSs) for progressive training. Based on NeuroAes-Data, we develop **AesMind**, an MLLM-based neuroaesthetic understanding and assessment model trained via a progressive three-stage paradigm, including skill harnessing for cognitive prompt initialization, supervised fine-tuning (SFT) for cross-modal knowledge internalization, and group relative policy optimization (GRPO) adapted for explicit reasoning elicitation. Extensive experiments demonstrate that AesMind achieves state-of-the-art performance across NeuroAes-Bench, substantially outperforming existing models on all continuous emotion dimensions. Furthermore, AesMind also exhibits remarkable cross-dataset generalization. Its ability to generate constructive photographic critiques and improvement strategies confirms an intrinsic grasp of aesthetic mecha-

nisms, elevating the model from merely identifying surface features to structurally decoding photographic art.

The main contributions are summarized as follows:

- We propose a **neuroaesthetics-grounded framework**, to establish a comprehensive photographic aesthetic experience taxonomy encompassing the *sensory-motor*, *emotion-valuation*, and *knowledge-meaning*.
- We construct **NeuroAes** dataset, comprising **NeuroAes-Bench** and **NeuroAes-Data**, which contains over 1.2M fine-grained neuroaesthetic labels and 150K mean opinion scores (MOSs) derived from 2.25M human ratings.
- We introduce **AesMind**, a progressive Skill-SFT-GRPO paradigm that combines cognitive initialization, knowledge internalization, and a novel continuous hybrid-reward optimization, to elicit explicit chain-of-aesthetics reasoning.
- Extensive experiments demonstrate AesMind's state-of-the-art performance on NeuroAes-Bench and robust cross-dataset generalization, validating its ability to transcend surface-level visual recognition and structurally decode photographic art through constructive aesthetic critiques.

2 Related Work

2.1 Neuroaesthetics and Aesthetic Emotion

Neuroaesthetics characterizes human aesthetic experience through the dynamic interaction of sensory-motor, emotion-valuation, and knowledge-meaning systems (Chatterjee & Vartanian, 2014; Leder et al., 2004). Within emotion-valuation, cognitive psychology strictly distinguishes general evoked emotions from specific aesthetic emotions, uniquely quantifying them via the valence-arousal-dominance (VAD) space (Menninghaus et al., 2019; Warriner et al., 2013). Recently, MLLMs have been increasingly applied to visual emotion understanding, yielding numerous benchmarks (Xie et al., 2024; Gao et al., 2026; Zhang et al., 2026; Lian et al., 2025). However, these works predominantly focus on general evoked emotions or coarse sentiment classification, thereby overlooking the unique characteristics of photographic aesthetic emotions. To address this, we operationalize the VAD space to distinctly quantify both the evoked emotions elicited by visual content and the aesthetic emotions induced by photographic execution.

2.2 Image Aesthetics Assessment

IAA has historically evolved from regressing scalar MOS on datasets like AVA (Murray et al., 2012) to exploring personalized aesthetic preferences (Ren et al., 2017; Yang et al., 2026b; Hayashi et al., 2026). Recognizing the limitations of isolated scores, recent efforts incorporate specific photographic attributes (Huang et al., 2024a) and shift towards hierarchical critique generation (Li et al., 2025a; Liu et al., 2026). Furthermore, pioneering studies advocate aligning computational models with human cognitive processes rather than relying on black-box mapping (He et al., 2022; Chen et al., 2025). Despite these advancements, existing evaluation dimensions remain largely heuristic. Constrained by the absence of a unified cognitive foundation, current models struggle to articulate a structured aesthetic reasoning process that connects sensory perception, emotional valuation, and semantic critique.

2.3 Multimodal Large Language Models for Aesthetics

The recent progress of MLLMs has advanced visual perception and IAA, yielding benchmarks and models, such as AesBench (Huang et al., 2024b), UNIAA (Zhou et al., 2024), and ArtiMuse (Cao et al., 2025), which combine numeric prediction with textual critique generation. Although research highlights the comparative reasoning capabilities of MLLMs for assessment (Kil et al., 2024; Teng et al., 2026), current aesthetic models rely predominantly on SFT methods (Hu et al., 2021). While SFT enables fluent text generation, it essentially memorizes mappings without endowing explicit reasoning trajectories. To bridge this gap, we adapt GRPO with a continuous hybrid reward mechanism, enabling the model to move beyond rote memorization and explicitly elicit an aesthetic chain-of-though reasoning process.

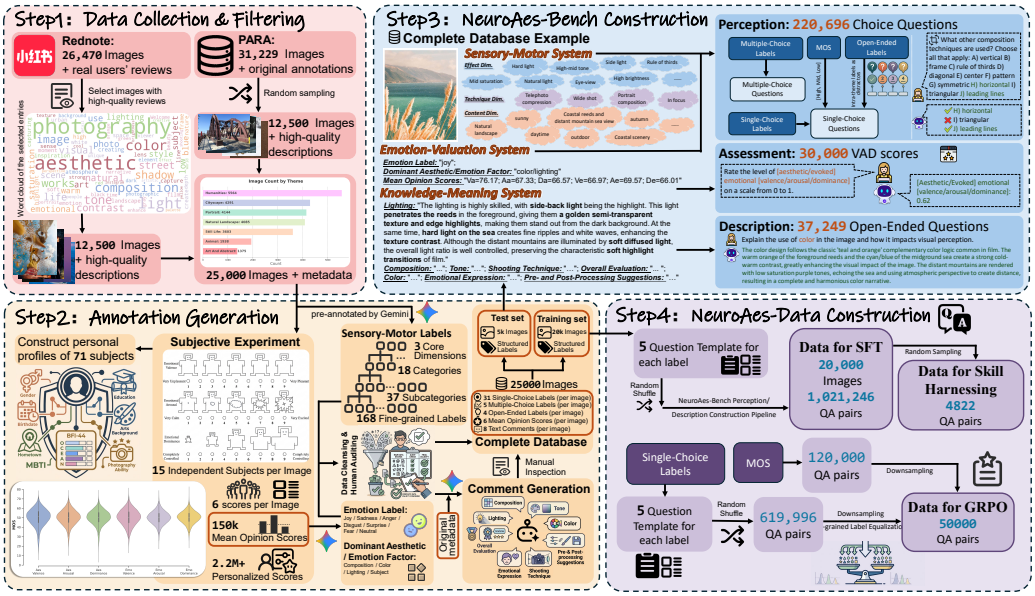


Figure 2: The construction pipeline of the NeuroAes dataset. The process comprises four steps, including (1) diverse image collection from Rednote (Xiaohongshu) and PARA, (2) human-AI collaborative annotation yielding fine-grained labels and over 150K subjective MOS ratings, (3) the formulation of the NeuroAes-Bench across three cognitive tasks, and (4) the systematic construction process of progressive training QA pairs in NeuroAes-Data tailored for the skill harnessing, SFT, and GRPO stages, respectively.

3 NeuroAes: A Neuroaesthetics-Grounded Framework and Benchmark

To operationalize our neuroaesthetic framework for computational aesthetic understanding, we construct **NeuroAes**, a comprehensive dataset comprising 25,000 images and over 2.25 million human annotations. We detail the computational taxonomy, data collection, subjective data processing, and the derived evaluation benchmark in Figure 2.

3.1 A Computational Taxonomy for Photographic Aesthetics

We design a specialized taxonomy for photographic aesthetics by functionally mapping three neuroaesthetic systems into computable dimensions:

(1) **Sensory-Motor System.** We map this system to the objective visual and semantic attributes of a photograph. We decompose it into *visual effects* (e.g., lighting, composition), *shooting parameters* (e.g., focal length, focus), and *semantic content* (e.g., subjects, events).

(2) **Emotion-Valuation System.** This dimension characterizes affective responses across discrete and continuous spectrums. First, we capture *evoked emotions* using six discrete Ekman categories alongside a neutral state. Second, we quantify both *aesthetic and evoked emotions* within a continuous valence-arousal-dominance (VAD) space, where aesthetic valence reflects subjective visual appeal, arousal characterizes immediate strikingness, and dominance reflects narrative clarity. Crucially, we explicitly annotate the *dominant influencing factors* to ground these subjective affective responses in concrete visual attributes.

(3) **Knowledge-Meaning System.** Integrating the preceding perceptual and affective information, this highest cognitive tier yields an 8-dimension professional photographic critique. This encompasses composition, tone, lighting, color, shooting technique, emotional expression, overall evaluation, and actionable pre&post-processing advice.

3.2 Data Collection and Subjective Experiment

We collect 25,000 photographic images from Xiaohongshu and PARA to ensure broad aesthetic diversity. Then Gemini-3-Pro (Google DeepMind, 2025) is used to generates initial sensory and

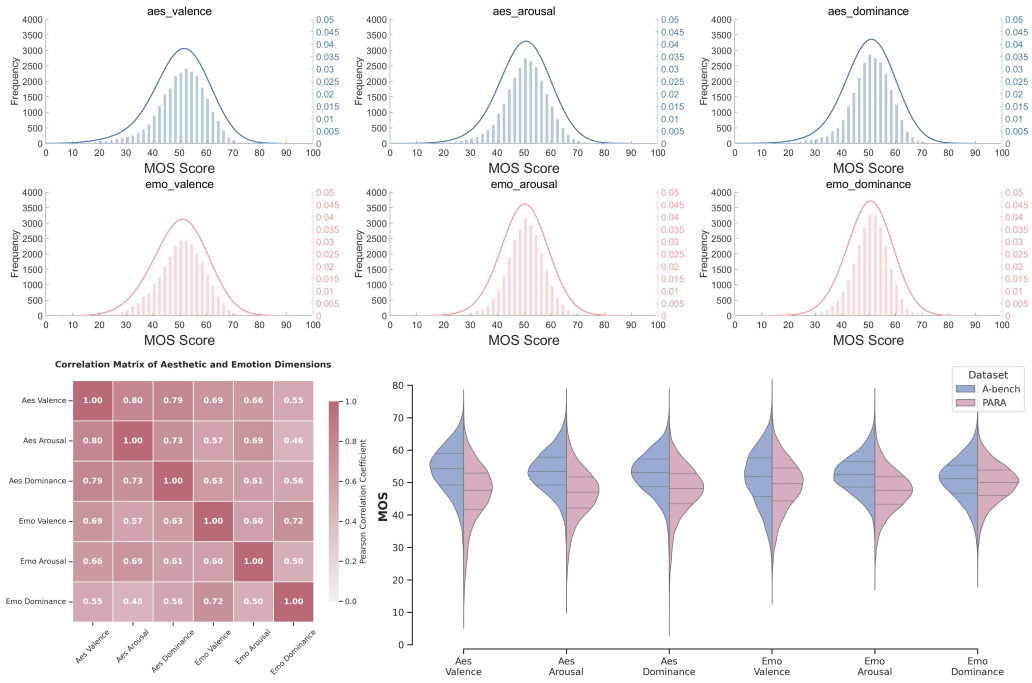


Figure 3: Statistical analysis of the subjective affective annotations. The VAD score distributions exhibit broad coverage across all aesthetic and evoked dimensions. The correlation matrix confirms these affective concepts are intrinsically related yet non-redundant. Furthermore, the violin plots demonstrate that the Xiaohongshu subset introduces a crucial high-aesthetic data density that is typically scarce in standard datasets like PARA.

knowledge annotations, which are subsequently verified and corrected by trained human annotators. For subjective affective evaluation, annotators rate each image on six continuous VAD dimensions using a 1–9 Likert-type Self-Assessment Manikin (SAM) scale. To prevent annotator-specific leakage, training and testing sets use strictly disjoint subject pools, with the 5,000 test images evaluated by a fixed group of 15 raters.

3.3 Subjective Data Processing and Analysis

Following the subjective data processing guidelines recommended by ITU, we standardize the ratings of each subject using a Z-score transformation to reduce annotator-specific bias, linearly rescale the standardized scores to $[0, 1]$, and average valid ratings to obtain the final MOSs as follows:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \qquad z'ij = \frac{zij + 3}{6}, \qquad \text{MOS}_j = \frac{1}{M} \sum_{i=1}^{M} z'_{ij}. \tag{1}$$

This yields 150,000 MOS annotations across six dimensions, which are further used to generate professional photographic critiques.

Figure 3 summarizes the distribution and correlation characteristics of the six affective dimensions.As illustrated in the figure, the six VAD scores exhibit Pearson correlations between 0.46 and 0.80, confirming these affective concepts are intrinsically related yet non-redundant. Morevover, the overall MOS distributions reveal broad aesthetic coverage of our NeuroAes. Notably, the Xiaohongshu subset provides a rich density of exceptional photography, mitigating the scarcity of highly aesthetic works in standard image quality datasets and facilitating expert-level learning.

3.4 Benchmark and Training Database Construction

NeuroAes-Bench evaluates three complementary tasks. **Perception** converts fine-grained annotations into single- and multiple-choice questions and uses the proposed Guessing-Penalty Accuracy

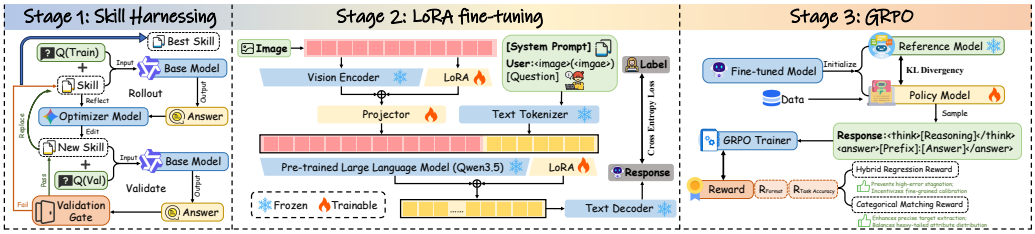


Figure 4: The progressive three-stage training paradigm of AesMind. Stage 1 employs skill harnessing to initialize meta-cognitive system prompts. Stage 2 internalizes cross-modal photographic knowledge via LoRA fine-tuning. Finally, Stage 3 elicits explicit chain-of-aesthetics reasoning through Group Relative Policy Optimization (GRPO), driven by a novel continuous hybrid regression reward and a categorical matching reward.

to reduce the effect of random guessing. For a single-choice question with N options,

$$S = \begin{cases} 1, & \text{if correct} \\ -\frac{1}{N-1}, & \text{if incorrect.} \end{cases} \tag{2}$$

Multiple-choice scores are computed from option-level accuracy and normalized relative to a random-guessing baseline. **Assessment** directly predicts the six continuous VAD scores and is evaluated using SRCC and PLCC against human MOS. **Description** evaluates generated critiques using LLM-based judgment, semantic similarity, and lexical overlap, including Qwen3.5-35B (Team, 2026), Gemini-3.1-Pro (Google DeepMind, 2025), SimCSE, Word2Vec, ROUGE-L, and METEOR. NeuroAes-Data is transformed into over one million QA pairs for progressive training, with task-specific subsets used for skill harnessing, SFT, and GRPO. Detailed construction pipeline is shown in Figure 2.

4 AesMind: The Cognitive Aesthetic MLLM

To operationalize our framework, we develop **AesMind**, a progressive three-stage multimodal architecture comprising meta-cognitive initialization, knowledge internalization, and explicit reasoning elicitation and the subsections below.

4.1 Skill Harnessing for Cognitive Initialization

To bridge the semantic gap between abstract aesthetic concepts and the model's linguistic space, we adopt skill harnessing (Yang et al., 2026a). Using a curated QA subset from NeuroAes-Data and Gemini-3.1-Pro as the teacher, we iteratively refine eight cognitive "skills" corresponding to our evaluation taxonomy. These optimized prompts guide all downstream training and inference, ensuring appropriate cognitive granularity.

4.2 SFT for Knowledge Internalization

Next, we internalize photographic knowledge by fine-tuning a Qwen3.5-9B backbone on over one million QA pairs. To prevent catastrophic forgetting, we employ Low-Rank Adaptation (LoRA) across the visual and textual branches while fully updating the vision-language merger. This Supervised Fine-Tuning (SFT) phase maps visual features to structural aesthetic labels and professional critiques.

4.3 GRPO for Explicit Reasoning Elicitation

To elicit explicit reasoning and refine continuous perception, we apply Group Relative Policy Optimization (GRPO) on a balanced 50K QA subset. We optimize task-specific behaviors via a KL-regularized dual-component reward system:

Format Reward. $R_{\text{fmt}}^{(i)}$ ensures the response encapsulates reasoning within `<think>` tags and conclusions within `<answer>` tags. Initialized with a negative baseline to prevent early gradient

Table 1: Performance comparisons on the perception task of NeuroAes-Bench. The best results and the second-best results are highlighted in **red** and **blue**, respectively.

Model	Motor System				Emotion System			Overall↑
	Effect ↑	Technical↑	Content↑	Overall ↑	Emotion↑	Aesthetic↑	Overall↑	
Random guess	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Medium-scale open-source MLLMs								
InternVL3-8B (Zhu et al., 2025)	22.60	25.29	52.76	31.77	14.68	16.52	15.64	29.09
InternVL3.5-8B (Wang et al., 2025)	28.20	32.33	48.03	34.35	15.03	16.30	15.74	30.39
Kimi-VL (Team et al., 2025)	27.03	29.91	56.62	35.71	21.97	20.10	**20.97**	32.79
MiniCPM-V-4.5 (Yu et al., 2025)	34.17	42.17	67.22	44.94	19.65	13.87	16.43	38.70
LLaVa-OneVision-1.5-8B (An et al., 2025)	31.64	40.03	66.73	41.70	**23.05**	15.53	18.81	35.80
Qwen2.5-VL-7B (Bai et al., 2025)	27.55	28.10	55.17	35.59	15.87	12.55	14.20	31.53
Qwen3-VL-8B (Yang et al., 2025)	30.34	42.23	52.25	37.90	17.50	15.20	16.22	33.47
Qwen3.5-9B (Team, 2026)	**42.79**	**53.57**	**69.80**	**49.59**	22.37	16.86	19.86	**42.05**
Large-scale open-source MLLMs								
Qwen3-VL-30B (Yang et al., 2025)	27.85	38.82	58.38	37.81	17.23	12.43	14.56	33.01
Qwen3.5-35B (Team, 2026)	34.47	50.45	66.94	45.41	17.34	**20.94**	19.30	40.07
Proprietary MLLMs								
Gemini-2.5-Pro (Comanici et al., 2025)	21.07	25.43	35.28	25.59	10.39	11.03	10.75	22.53
Gemini-3-Pro (Google DeepMind, 2025)	25.81	30.30	39.79	30.29	11.04	10.79	10.90	26.28
MLLMs for Image Aesthetics Assessment								
Artimuse (Cao et al., 2025)	27.91	27.76	68.30	38.85	12.94	14.28	13.71	33.11
AesMind (Ours)	**66.57**	**80.80**	**76.09**	**70.89**	**43.22**	**44.42**	**43.75**	**65.36**

Table 2: Performance comparisons on the assessment task of NeuroAes-Bench. Metrics are SRCC↑/PLCC↑. ⋆ denotes models that have been fine-tuned on our NeuroAes-Data. The best results and the second-best results are highlighted in **red** and **blue**, respectively.

Model	Aesthetic Emotion			Evoked Emotion			Overall↑
	Valence	Arousal	Dominance	Valence	Arousal	Dominance	
Medium-scale open-source MLLMs							
InternVL3-8B (Zhu et al., 2025)	0.5424 / 0.5342	0.4596 / 0.4736	0.1488 / 0.3623	0.5523 / 0.5340	0.2273 / 0.2260	0.0429 / 0.0353	0.3289 / 0.3609
InternVL3.5-8B (Wang et al., 2025)	0.4680 / 0.4153	0.3928 / 0.4019	0.2600 / 0.3055	0.4964 / 0.4761	0.1455 / 0.1541	-0.1651 / -0.1573	0.2663 / 0.2659
Kimi-VL (Team et al., 2025)	0.6241 / 0.6420	0.5423 / 0.5849	0.4609 / 0.5039	0.6314 / 0.6208	0.3949 / 0.4197	-0.0062 / 0.0066	0.4412 / 0.4630
LLaVA-OneVision-1.5-8B (An et al., 2025)	0.5090 / 0.5459	0.4462 / 0.5192	0.2976 / 0.4744	0.5321 / 0.5271	0.3301 / 0.3683	-0.0896 / -0.0888	0.3376 / 0.3910
Qwen2.5-VL-7B (Bai et al., 2025)	0.5975 / 0.5691	0.5301 / 0.5265	0.3117 / 0.4069	0.6208 / 0.5800	0.2921 / 0.2975	-0.0076 / -0.0291	0.3908 / 0.3918
Qwen3-VL-8B (Yang et al., 2025)	0.0186 / 0.0153	0.0043 / -0.0008	0.0071 / 0.0117	0.0091 / 0.0213	0.0158 / 0.0072	0.0162 / 0.0359	0.0118 / 0.0151
Qwen3.5-9B (Team, 2026)	0.4336 / 0.4543	0.4072 / 0.4359	0.0435 / 0.0407	0.2705 / 0.2845	0.3386 / 0.3323	0.0285 / 0.0203	0.2537 / 0.2613
Large-scale open-source MLLMs							
Qwen3-VL-30B (Yang et al., 2025)	0.5321 / 0.6109	0.5144 / 0.5759	0.2503 / 0.4025	0.5711 / 0.5709	0.3627 / 0.3572	-0.143 / -0.109	0.3479 / 0.4014
Qwen3.5-35B (Team, 2026)	0.1453 / 0.1446	0.1525 / 0.1673	0.1926 / 0.2126	0.1635 / 0.1583	-0.1818 / -0.1892	0.0559 / 0.0476	0.0880 / 0.0902
MLLMs for Image Aesthetics Assessment							
MUSIQ (Ke et al., 2021) ⋆	0.5478 / 0.5956	0.5185 / 0.5724	0.4738 / 0.5142	0.6498 / **0.6731**	**0.4612 / 0.4813**	**0.5608 / 0.5886**	**0.5353 / 0.5709**
ArtiMuse (Cao et al., 2025)	0.2907 / 0.4255	0.2546 / 0.367	0.2773 / 0.3647	0.3570 / 0.3535	0.2293 / 0.2517	-0.0685 / -0.0614	0.2234 / 0.2835
ArtiMuse (Cao et al., 2025) ⋆	**0.6884 / 0.7095**	**0.5898 / 0.6321**	**0.5477 / 0.5939**	**0.6552** / 0.6413	0.3655 / 0.4102	0.3528 / 0.3553	0.5332 / 0.5571
AesMind (Ours) ⋆	**0.7433 / 0.7779**	**0.6832 / 0.7287**	**0.6783 / 0.7305**	**0.7789 / 0.7790**	**0.6318 / 0.6609**	**0.6762 / 0.6799**	**0.6986 / 0.7261**

vanishing, it grants additive rewards for correct tag presence, ordering, and prefix matching, while penalizing repetition and empty blocks.

Hybrid Regression Reward. Standard threshold or ℓ_1-norm penalties struggle with continuous assessment. We engineer a hybrid regression reward $R_{\text{reg}}^{(i)}$ balancing global stability with local precision:

$$R_{\text{reg}}^{(i)} = \lambda_{\text{base}} \cdot \max(0, 1 - \delta) + \lambda_{\text{peak}} \cdot \exp\left(-\frac{\delta^2}{2\sigma^2}\right), \tag{3}$$

where $\delta = |s - \hat{s}_i|$. The linear *guidance zone* (λ_{base}) prevents high-error stagnation, while the Gaussian *high-precision zone* ($\lambda_{\text{peak}}, \sigma$) incentivizes fine-grained calibration.

Categorical Matching Reward. For discrete perception, strict text-matching yields a high positive reward for correct target extraction. We incorporate a prior-based logarithmic adjustment for heavy-tailed aesthetic attributes:

$$R_{\text{cat}}^{(i)} = \begin{cases} \lambda_{\text{match}} + \min(-\alpha \ln(p_c + \epsilon), \gamma), & \text{if correct} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where λ_{match} is the base reward, p_c the category prior, α the long-tail scale, and γ the maximum limit.The final reward r_i integrates these components:

$$r_i = R_{\text{fmt}}^{(i)} + \mathbb{1}_{\text{reg}} \cdot R_{\text{reg}}^{(i)} + \mathbb{1}_{\text{cat}} \cdot R_{\text{cat}}^{(i)}, \tag{5}$$

where indicator functions $\mathbb{1}_{\text{reg}}$ and $\mathbb{1}_{\text{cat}}$ act as curriculum gates. They activate content rewards only if $R_{\text{fmt}}^{(i)}$ exceeds a dynamic threshold τ, compelling structural mastery before semantic optimization.

Table 3: **Quantitative results on the NeuroAes-Bench description task.** We evaluate model performance comprehensively across three paradigms: LLM-based evaluation, semantic embedding similarity, and traditional lexical matching. The best results and the second-best results are highlighted in **red** and **blue**, respectively.

Model	LLM-as-a-Judge		Semantic		Lexical	
	Qwen3.5	Gemini-3.1	SimCSE	Word2Vec	ROUGE-L	METEOR
InternVL3.5-8B (Wang et al., 2025)	0.7618	0.8126	0.8252	0.9043	0.1604	0.1997
Kimi-VL (Team et al., 2025)	0.7132	0.7403	0.8053	0.8921	0.1852	0.1503
LLaVa-OneVision-1.5-8B (An et al., 2025)	**0.9499**	1.0201	0.8087	0.8954	**0.1895**	0.1635
Qwen3-VL-8B (Yang et al., 2025)	0.6549	0.7279	0.8427	0.9146	0.1301	0.1934
Qwen3.5-9B (Team, 2026)	0.6520	0.7689	0.8285	0.9035	0.1305	0.1750
Qwen3.5-35B (Team, 2026)	0.6701	0.7894	0.8429	0.9109	0.1113	0.1920
Gemini-3-Pro (Google DeepMind, 2025)	0.8619	1.0088	**0.8637**	**0.9236**	0.1516	**0.2115**
Artimuse (Cao et al., 2025)	0.9310	**1.0319**	0.8245	0.9069	0.1688	0.1961
AesMind (Ours)	**1.1043**	**1.1840**	**0.8582**	**0.9262**	**0.2559**	**0.2335**

Table 4: **Cross-dataset generalization performance** on five conventional image aesthetics benchmarks. † Results are taken directly from original papers as these models cannot be tested. ⋆ Results are trained only on AVA to compare the generalization ability. The best results and the second-best results are highlighted in **red** and **blue**, respectively.

Model	AVA		PARA		TAD66K		FLICKR-AES		ArtiMuse-10K	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Traditional Models										
MUSIQ (Ke et al., 2021)	0.225	0.258	0.490	0.600	0.099	0.149	0.150	0.216	-0.060	-0.074
TANet (He et al., 2022) †	0.758	0.765	–	–	**0.513**	**0.531**	–	–	–	–
VILA (Lin et al., 2024)	0.776	0.775	0.651	0.658	0.418	0.444	0.616	0.645	0.273	0.268
AesMamba (Gao et al., 2024) †	0.774	0.769	**0.936**	0.902	0.511	0.483	–	–	–	–
MLLMs for General-Purpose Applications										
Qwen2.5-VL-7B (Bai et al., 2025)	0.391	0.371	0.721	0.743	0.240	0.242	0.621	0.578	0.256	0.179
InternVL3-8B (Zhu et al., 2025)	0.364	0.332	0.667	0.693	0.203	0.191	0.553	0.459	0.187	0.157
MLLMs for Image Aesthetics Assessment										
Q-Align (Wu et al., 2024)	0.822	0.817	0.913	0.888	0.501	0.531	0.798	0.818	0.551	0.573
UniAA-LLaVA (Zhou et al., 2024) †	0.713	0.704	0.864	0.895	0.411	0.425	0.724	0.751	–	–
Next Token Is Enough (Li et al., 2025b) †	**0.828**	0.825	–	–	0.413	0.444	–	–	–	–
ArtiMuse (Cao et al., 2025)	0.827	**0.826**	**0.936**	**0.958**	0.510	**0.543**	**0.814**	**0.837**	**0.614**	**0.627**
AesMind (Ours)	**0.836**	**0.832**	**0.931**	**0.956**	**0.515**	**0.543**	**0.821**	**0.840**	**0.616**	**0.635**
Comparison of Generalization Ability										
Q-Align (Wu et al., 2024) ⋆	0.822	0.817	0.694	0.711	0.417	0.445	0.643	0.664	0.337	0.320
ArtiMuse (Cao et al., 2025) ⋆	**0.827**	**0.826**	**0.697**	**0.725**	**0.419**	**0.451**	**0.647**	**0.676**	**0.395**	**0.376**
AesMind (Ours) ⋆	**0.836**	**0.832**	**0.705**	**0.743**	**0.442**	**0.473**	**0.693**	**0.720**	**0.405**	**0.415**

5 Experiments

5.1 Experimental Setup

We implement AesMind using Qwen3.5-9B (Team, 2026) with image resolutions constrained between 262,144 and 393,216 pixels. Post-initialization, the model undergoes SFT via LoRA (rank 64, scaling factor 128) on both branches, utilizing a peak learning rate of 2×10^{-5} alongside a fully updated merger. The GRPO phase employs a generation group size of 4, a KL penalty of 0.04, and reduced learning rates of 1×10^{-6} for the language model and 5×10^{-7} for vision components. Reward hyperparameters are $\lambda_{base} = 0.4$, $\lambda_{peak} = 0.6$, and $\sigma = 0.05$. We evaluate zero-shot performance against generalist (e.g., LLaVA-OneVision, Gemini-Pro) and specialized (e.g., Q-Align (Wu et al., 2024), ArtiMuse (Cao et al., 2025)) models under strictly identical protocols.

5.2 Evaluation on NeuroAes-Bench

Perception. Table 1 shows that while generalist models stagnate between 10% and 20% on emotion system tasks, AesMind achieves a commanding overall accuracy of 65.36%—outperforming its Qwen3.5-9B base (42.05%) by over 23 absolute points. This validates our cognitive initialization and fine-grained taxonomy.

Assessment. Standard multimodal models fail catastrophically at continuous subjective regression (Table 2), frequently yielding near-zero correlations on dominance dimensions. Empowered by the

Table 5: Performance comparison on CogIP-Bench. **Bold** indicates the best performance. ⋆ denotes that the Spearman correlation is statistically insignificant.

Model	Aesthetics			Funniness			Emotional			Memorability		
	MSE↓	MAE↓	Spearman↑	MSE↓	MAE↓	Spearman↑	MSE↓	MAE↓	Spearman↑	MSE↓	MAE↓	Spearman↑
ArtiMuse (Cao et al., 2025)	9.549	2.550	0.048 ⋆	9.243	2.440	0.484	3.762	1.457	0.553	0.133	0.289	**0.058** ⋆
AesMind (Ours)	**3.646**	**1.470**	**0.595**	**2.961**	**1.346**	**0.667**	**2.001**	**1.041**	**0.595**	**0.089**	**0.251**	0.016 ⋆

Table 6: Ablation studies on training stages and GRPO tasks. **Bold** indicates the best performance, and underline indicates the second best.

	Training Strategy			Perception			Assessment		Description	
	Stage 1	Stage 2	Stage 3	Motor	Emotion	Overall	Overall		LLM	Semantic
No.	Skill	LoRA	GRPO	ACC↑	ACC↑	ACC↑	SRCC↑	PLCC↑	Gemini-3.1↑	Word2Vec↑
(1)				53.21	19.28	45.65	0.2537	0.2613	0.7689	0.9035
(2)		✓		70.07	31.64	62.23	0.4871	0.5270	1.0513	0.9180
(3)			✓	48.81	23.44	43.56	0.3738	0.3816	0.8527	0.8914
(4)		✓	✓	71.71	44.73	66.21	0.6532	0.6707	1.2083	0.9255
(5)	✓			52.74	24.36	46.95	0.4492	0.4956	**1.2328**	0.9111
(6)	✓	✓		**76.31**	**47.95**	**70.53**	0.5096	0.5420	1.1685	0.9250
(7)	✓	✓	✓	70.89	43.75	65.36	**0.6986**	**0.7261**	1.1840	**0.9262**

continuous hybrid reward, AesMind establishes a new state-of-the-art across all six dimensions, achieving an overall SRCC of 0.6986 and PLCC of 0.7261.

Description. Evaluated under the LLM-as-a-Judge protocol (Table 3), AesMind scores 1.1043 and 1.1840, vastly outperforming baselines in structured critique generation. It additionally achieves the highest lexical alignment scores, demonstrating robust semantic gap bridging during SFT.

5.3 Cross-Dataset Evaluation

Table 4 verifies broad generalization on five conventional aesthetic benchmarks (e.g., AVA, PARA). AesMind consistently outperforms leading specialized models like Q-Align and ArtiMuse under strict generalization settings. Furthermore, evaluations on CogIP-Bench (Table 5) reveal drastic Mean Squared Error reductions and rank correlation improvements across Funniness and Emotional perception. This confirms that decoupling aesthetic appeal from evoked emotions yields universal, rather than dataset-specific, affective representations.

5.4 Ablation Study

Table 6 ablates our progressive training components. Supervised Fine-Tuning establishes the essential structural baseline, acting as a mandatory prerequisite for effective Group Relative Policy Optimization. Building upon SFT, the dual-reward GRPO drives subjective alignment, significantly improving continuous assessment correlations. Concurrently, skill initialization enhances cognitive granularity across perception and description tasks. Ultimately, the complete AesMind architecture achieves highly comprehensive capabilities, maintaining robust discrete perception while delivering strictly optimal performance on the primary continuous assessment task.

6 Conclusion

We presented a comprehensive computational framework grounded in neuroaesthetic principles to advance multimodal aesthetic reasoning. **NeuroAes** provides a large-scale benchmark that operationalizes the aesthetic cognitive chain by explicitly decoupling aesthetic appeal from evoked emotions. Building upon this taxonomy, we developed **AesMind**, an MLLM trained via a progressive three-stage paradigm. By integrating meta-cognitive skill harnessing, structural SFT, and GRPO featuring a novel continuous hybrid regression reward, AesMind effectively bridges the semantic gap between objective visual features and complex human subjectivity. Extensive evaluations demonstrate state-of-the-art performance across diverse perception and continuous assessment tasks, establishing a robust foundation for future subjective multimodal alignment.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets and for aiding and polishing the writing of this manuscript. Specifically, AI models were employed to construct the fine-grained aesthetic annotations and structural textual critiques for the NeuroAes dataset based on human-derived Mean Opinion Scores (MOS). We have not used generative AI tools for formulating the core theoretical claims, developing the primary algorithmic framework, or creating the empirical methodology.

We have thoroughly reviewed all AI-assisted work. The AI-generated dataset annotations were meticulously cross-validated against human perceptual ground truths to ensure factual alignment and prevent hallucination. Furthermore, all AI-polished textual segments were carefully inspected and edited by the authors to guarantee scientific accuracy, maintain the original academic intent, and prevent potential plagiarism. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research introduces NeuroAes, a comprehensive dataset comprising photographic images alongside multidimensional aesthetic and emotional annotations. The images utilized in our benchmark are sourced from publicly accessible repositories, and we have taken diligent care to ensure that no personally identifiable information or offensive content is inappropriately exposed. The human-derived Mean Opinion Scores (MOS) acting as the foundation for our dataset were collected following standard ethical guidelines for crowdsourcing, ensuring annotator anonymity and fair compensation.

Furthermore, we recognize that image aesthetics and emotional valuation are inherently subjective, culturally situated, and highly diverse. While the AesMind framework demonstrates advanced capabilities in aesthetic critique, we strongly caution against deploying such models as absolute arbiters of artistic value. Unregulated application could inadvertently marginalize minority aesthetic expressions or reinforce dominant cultural biases. We encourage future research to continuously explore culturally diverse and personalized aesthetic alignment.

REPRODUCIBILITY STATEMENT

We are highly committed to ensuring the full reproducibility of the AesMind framework and the NeuroAes-Bench evaluation. To facilitate this, the complete hyperparameter configurations for all progressive training stages (Skill, SFT, and GRPO), along with the exact system prompts, categorical taxonomies, and reward formulation logic, are exhaustively documented in the Appendices. Furthermore, anonymized source code for the training pipeline and evaluation scripts, alongside a sample subset of the NeuroAes dataset, are provided in the supplementary materials. Upon acceptance, the complete NeuroAes dataset, the NeuroAes-Bench evaluation suite, and all pre-trained AesMind model weights will be fully open-sourced to the community.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

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. *arXiv preprint arXiv:2507.14533*, 2025.

Anjan Chatterjee and Oshin Vartanian. Neuroaesthetics. *Trends in Cognitive Sciences*, 18(7):370–375, 2014. doi: 10.1016/j.tics.2014.03.003.

Yiming Chen, Junlin Han, Tianyi Bai, Shengbang Tong, Filippos Kokkinos, and Philip Torr. From pixels to feelings: Aligning mllms with human cognitive perception of images, 2025. URL https://arxiv.org/abs/2511.22805.

Gheorghe Comanici, Eric Bieber, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Fei Gao, Yuhao Lin, Jiaqi Shi, Maoying Qiao, and Nannan Wang. Aesmamba: Universal image aesthetic assessment with state space models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 7444–7453, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681011. URL https://doi.org/10.1145/3664647.3681011.

Lancheng Gao, Ziheng Jia, Zixuan Xing, Wei Sun, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Eemo-logic: A unified dataset and multi-stage framework for comprehensive image-evoked emotion assessment, 2026. URL https://arxiv.org/abs/2602.01173.

Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, November 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Takato Hayashi, Hiroaki Takahara, Candy Olivia Mawalim, Hiromi Narimatsu, Akisato Kimura, Shiro Kumano, and Shogo Okada. Xpass-vis: A dataset for cross-domain personalized image aesthetic assessment, 2026. URL https://arxiv.org/abs/2606.15629.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/132. URL https://doi.org/10.24963/ijcai.2022/132. Main Track.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Yipo Huang, Leida Li, Pengfei Chen, Jinjian Wu, Yuzhe Yang, Yaqian Li, and Guangming Shi. Coarse-to-fine image aesthetics assessment with dynamic attribute selection. *IEEE Transactions on Multimedia*, 26:9316–9329, 2024a. doi: 10.1109/TMM.2024.3389452.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Leida Li, Pengfei Chen, Yu Liu, Weisi Lin, and Guangming Shi. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28798–28827. Curran Associates, Inc., 2024. doi: 10.52202/079017-0906. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/32923dff09f75cf1974c145764a523e2-Paper-Datasets_and_Benchmarks_Track.pdf.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation, 2016. URL https://arxiv.org/abs/1606.01621.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95:489–508, 2004. doi: 10.1348/0007126042369811.

Leida Li, Xiangfei Sheng, Pengfei Chen, Jinjian Wu, and Weisheng Dong. Towards explainable image aesthetics assessment with attribute-oriented critiques generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1464–1477, 2025a. doi: 10.1109/TCSVT.2024.3470870.

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, 2025b. URL https://arxiv.org/abs/2503.06141.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. Ov-mer: Towards open-vocabulary multimodal emotion recognition, 2025. URL https://arxiv.org/abs/2410.01495.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. URL https://arxiv.org/abs/2312.07533.

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, 2026. URL https://arxiv.org/abs/2512.23413.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Ines Schindler, Julian Hanich, Thomas Jacobsen, and Stefan Koelsch. What are aesthetic emotions? *Psychological Review*, 126(2):171–195, 2019. doi: 10.1037/rev0000135.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, 2012. doi: 10.1109/CVPR.2012.6247954.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomír Mech, and David J. Foran. Personalized image aesthetics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 638–647, 2017. doi: 10.1109/ICCV.2017.76.

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

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Jiajie Teng, Huiyu Duan, Sijing Wu, Jiarui Wang, Xilei Zhu, Jianing Jin, Wei Shen, Xiongkuo Min, and Guangtao Zhai. MIMIC-bench: Exploring the user-like thinking and mimicking capabilities of multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=J7wc4G6woS.

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

A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45:1191–1207, 2013. doi: 10.3758/s13428-012-0314-x. URL https://doi.org/10.3758/s13428-012-0314-x.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. *arXiv preprint arXiv:2404.16670*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. Skillopt: Executive strategy for self-evolving agent skills, 2026a. URL https://arxiv.org/abs/2605.23904.

Zhichao Yang, Tianjiao Gu, Zhixianhe Zhang, Xiangfei Sheng, Pengfei Chen, and Leida Li. Personalized image aesthetic assessment via preference-rich sample mining and cohort merging, 2026b. URL https://arxiv.org/abs/2607.15752.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models, 2026. URL https://arxiv.org/abs/2508.09210.

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, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX

A OVERVIEW

This supplementary material provides comprehensive details regarding the theoretical foundations, dataset construction, training methodology, and extended results of the NeuroAes framework and AesMind model. The content is organized as follows: Sec. B details the neuroaesthetic foundations, the operationalization of aesthetic VAD, and the derivation of our three-system cognitive taxonomy. Sec. C systematically compares existing image aesthetics datasets with NeuroAes. Sec. D delineates the data curation process, subjective experiment protocols, and the prompt-driven human-AI collaborative annotation pipeline. Sec. E explains the formulation of NeuroAes-Bench and NeuroAes-Data, covering QA construction, evaluation criteria, and specific data sampling strategies for SFT and GRPO. Sec. F elaborates on the algorithmic and implementation specifics of skill harnessing, SFT, and GRPO, including reward designs, inference procedures, and comprehensive hyperparameters. Finally, Sec. G presents extended ablation studies, fine-grained quantitative analyses, and qualitative visualizations demonstrating real-world generalization.

B THEORETICAL FOUNDATIONS AND COGNITIVE TAXONOMY

B.1 NEUROAESTHETICS FOUNDATIONS

The NeuroAes framework is profoundly anchored in the neuroaesthetics theory proposed by Chatterjee and Vartanian (2014), which conceptualizes aesthetic experiences not as monolithic phenomena, but as emergent states arising from the interactions of three distinct neural systems: the sensory-motor, emotion-valuation, and knowledge-meaning systems.

According to this "aesthetic triad", the *sensory-motor system* handles the automatic processing of elemental visual features and perceptual mechanics. The *emotion-valuation system* governs the affective and rewarding aspects of the aesthetic encounter. Finally, the *knowledge-meaning system* contextualizes the experience through cultural background, expertise, and semantic interpretation.

Crucially, in the context of the NeuroAes framework, we adopt this triad strictly as a functional and computational organizational blueprint. Rather than claiming to directly measure or simulate biological neural activity, we utilize these three interacting systems to systematically decouple and structure the multimodal cognitive processes inherent in photographic aesthetic assessment. This allows our computational model to progress logically from low-level visual perception, through intermediate affective valuation, to high-level professional critique.

B.2 AESTHETIC EMOTION

To comprehensively model the affective dimension of aesthetic experiences, NeuroAes integrates the theoretical framework of "aesthetic emotions" proposed by Menninghaus et al. (2019). This work draws a fundamental distinction between aesthetic and everyday emotions. Whereas everyday emotions are typically triggered by practical relevance, goal-conduciveness, or semantic reality, aesthetic emotions arise specifically from the evaluation of an object's sensory, formal, and stylistic qualities. They are directed at the artistic execution and expressive power of a stimulus, ultimately culminating in states of aesthetic pleasure or displeasure.

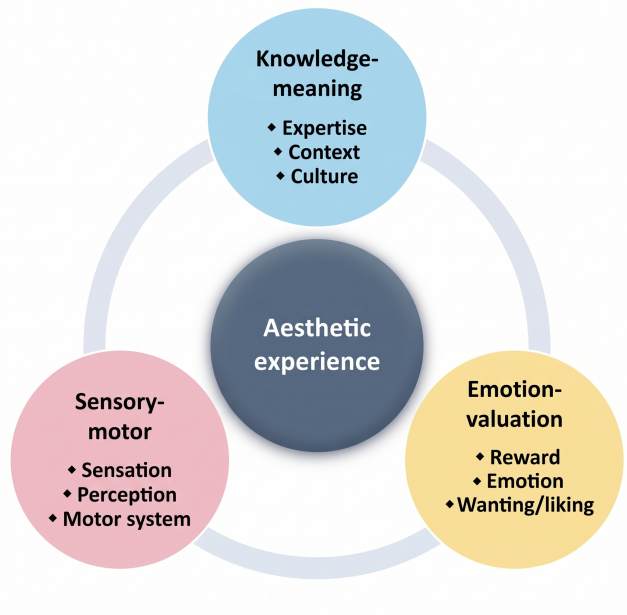


Figure 5: **The aesthetic triad.** Aesthetic experiences emerge from the complex interactions between sensory-motor, emotion-valuation, and knowledge-meaning systems. Adapted from Chatterjee and Vartanian (2014). In NeuroAes, this biological framework serves as the structural foundation for our computational cognitive taxonomy.

A core tenet of this theory is that aesthetic emotions stem from cognitive appraisals focused on the *form and execution* of an object rather than merely its *content*. Consequently, individuals can experience profound aesthetic pleasure even when exposed to stimuli depicting negative or distressing subject matter that would otherwise evoke aversive everyday emotions.

This theoretical distinction is critical for photographic aesthetic understanding. For instance, a masterfully composed photograph of a tragic event may elicit strong negative feelings (low evoked valence) while simultaneously inspiring deep aesthetic appreciation (high aesthetic valence). By explicitly decoupling aesthetic and evoked emotions within the Emotion-Valuation system, NeuroAes effectively disentangles the affective impact of semantic content from the evaluation of photographic execution. This dual-track modeling enables the framework to mitigate content-induced bias and generate highly nuanced, objective aesthetic assessments.

B.3 Operationalization of Aesthetic VAD

The Valence-Arousal-Dominance (VAD) model is a well-established continuous dimensional framework in affective computing and subjective evaluation, traditionally utilized to quantify psychological states induced by external stimuli (i.e., evoked emotions). Given that aesthetic appreciation inherently constitutes an affective response—as established by the aesthetic emotion framework (Sec. B.2)—we extend the continuous VAD space to explicitly quantify aesthetic emotions.

Crucially, while conventional evoked VAD measures the viewer's psychological reaction to the semantic *content* of an image, our proposed aesthetic VAD operationalizes the viewer's affective response specifically directed at the *photographic execution and artistic form*. The specific operational definitions designed for the NeuroAes framework are detailed in Table 7.

To ensure consistent and reliable human annotation during the subjective experiments, these operational definitions were communicated to participants using explicit descriptive guidelines and visual anchors. For example, annotators were instructed that a landscape photograph with exceptional lighting and harmonious composition, yet depicting a tranquil scene, should receive a high Aesthetic Valence score but a moderate-to-low Aesthetic Arousal score. Conversely, a highly saturated, visually aggressive macro shot might score high on Aesthetic Arousal due to its immediate

Table 7: Operational definitions of the three aesthetic VAD dimensions in NeuroAes.

Dimension	Operational Definition
Aesthetic Valence	**Degree of visual appeal and aesthetic pleasure.** High scores indicate a highly pleasing, beautiful, or harmonious photographic execution. Low scores indicate an unappealing, visually displeasing, or poorly executed image.
Aesthetic Arousal	**Intensity of visual impact and strikingness.** High scores represent a breathtaking, highly striking, or visually arresting image upon initial perception. Low scores indicate a mundane, visually flat, or unstimulating execution.
Aesthetic Dominance	**Clarity of photographic intent and visual narrative.** High scores denote a clear, well-structured visual design where the subject or photographic intent is immediately apparent and effectively controlled. Low scores reflect an ambiguous, cluttered, or poorly articulated visual message.

visual impact, but could receive a lower Aesthetic Dominance score if the core subject remains difficult to recognize. By establishing these precise operational boundaries, we effectively minimize the ambiguity typically associated with subjective aesthetic scoring.

B.4 Construction of the Cognitive Taxonomy

To bridge the theoretical neuroaesthetic foundations with a computable, data-driven framework, we iteratively developed the NeuroAes cognitive taxonomy through extensive consultations with professional photographers and cognitive scientists. The primary objective was to translate the biological "aesthetic triad" into a structured, hierarchical annotation schema tailored specifically for photographic assessment.

The resulting taxonomy strictly follows a functional cognitive progression: from objective perception to subjective valuation, and finally to semantic interpretation. First, the **Sensory-Motor System** focuses on the objective visual and physical attributes of the photograph. To comprehensively capture these properties, we deconstruct this system into three facets: *Effect Evaluation* (e.g., lighting quality, tonal distribution, color harmony), *Framing & Optics* (e.g., camera setup, focal length, focus), and *Visual Semantics* (e.g., subjects, context, events). Second, the **Emotion-Valuation System** quantifies the subjective affective response. Crucially, it explicitly decouples the aesthetic VAD (evaluating the artistic execution) from the evoked VAD and discrete Ekman emotions (evaluating the semantic content), while also identifying the dominant photographic factors driving these responses. Finally, the **Knowledge-Meaning System** represents the highest level of cognitive interpretation. We formulated this system as an 8-dimensional professional photographic critique, mimicking how an expert photographer would verbally articulate the aesthetic merits and flaws of an image.

The complete hierarchical structure, encompassing single-choice, multiple-choice, open-ended labels, continuous VAD scores, and textual critiques, is exhaustively detailed in Table 8. This taxonomy guarantees that the dataset covers the full spectrum of photographic aesthetic comprehension, ensuring full reproducibility for subsequent model training.

C Comparison with Existing Aesthetics Datasets

As summarized in Table 9, NeuroAes is designed to address the limitations of sparse and monolithic annotations in existing Image Aesthetics Assessment (IAA) datasets. While NeuroAes is smaller in overall scale compared to massive consensus-driven datasets like AVA, and is strictly constrained to the photographic domain—excluding paintings or mixed visual arts (e.g., TAD66K, ArtiMuse-10K)—it deliberately trades sheer volume for annotation granularity. Unlike conventional datasets that primarily rely on global scalar scores (e.g., KonIQ-10K) or extremely limited attribute tags (e.g., 0–4 labels in AVA and PARA), NeuroAes provides 40 fine-grained labels per image. This

Table 8: Complete definitions and label spaces of the NeuroAes cognitive taxonomy.

System	Category	Dimension	Label Space / Specifics
Sensory-Motor	Effect	Effect Evaluation	Dynamics, Lighting, Composition, Color (outstanding / medium / poor).
		Shooting Angle	Camera View (top-down / low-angle / eye-level); Portrait Angle (side / three-quarter / front / back / uncertain); Portrait Composition (full-body / half-body / three-quarter body / bust / close-up / facial close-up / body-part close-up / back view / uncertain).
		Motion State	Dynamic State (dynamic-frozen / shallow-motion-blur / static); Dynamic Action (open-ended description).
		Light	Elements (open-ended description); Source (artificial / natural / uncertain / mixed); Direction (front / back / side / top / bottom / complex); Quality (soft / hard).
		Color	Temperature (cool tone / warm tone / black and white / mixed cool-warm); Primary Schemes (red, orange, yellow, green, cyan, blue, purple, gray); Saturation (low / medium / high); Brightness Level (low / medium / high); Color Usage (monochrome / analogous / complementary / mixed / black and white).
		Exposure	Overexposed / underexposed / properly exposed.
		Tone	Full tonal range / high-long tone / mid-long tone / low-long tone / high-mid tone / mid-mid tone / low-mid tone / high-short tone / mid short-tone / low-short tone.
		Image Quality	Artistic low quality / quality defect / normal.
		Composition	Primary & Secondary Composition (rule of thirds / center / horizontal / symmetric / diagonal / vertical / triangular / pattern / frame / leading lines / none).
	Shooting	Shot Scale	Extreme wide shot / wide shot / medium shot / close-up / close-up detail / extreme close-up / selfie / uncertain.
		Focal Length	Wide-angle / standard / telephoto compression / macro / uncertain.
		Comp. Orientation	Landscape / portrait / square.
		Focus	Out of focus / in focus.
	Semantics	Theme	Humanities / cityscape / natural landscape / portrait / still life / animal / art and abstract.
		Who	Number (single / two / multiple / none); Gender (male / female / uncertain); Age Group (infant / child / youth / adult / middle-aged / elderly / uncertain); Animals Present (people and animals / animals only / none).
		When	Weather (sunny / cloudy / rainy / snowy / foggy / uncertain); Time (early morning / daytime / evening / night / uncertain); Season (spring / summer / autumn / winter / uncertain).
		Where	Outdoor / indoor / indoor-outdoor mixed / inside vehicle / uncertain.
		What	Main Subject (open-ended phrase); Event (open-ended description).
Emotion-Valuation	Aesthetic Emo	Valence	Continuous MOS.
		Arousal	Continuous MOS.
		Dominance	Continuous MOS.
		Dominant Aes. Factor	Composition / lighting / color / subject.
	Evoked Emo	Valence	Continuous MOS.
		Arousal	Continuous MOS.
		Dominance	Continuous MOS.
		Emotion Label	Joy / anger / disgust / sadness / surprise / fear / neutral.
		Dominant Emo. Factor	Composition / lighting / color / subject.
Knowledge-Meaning	*Prof. Critique*	Composition Design Tonal Quality Lighting Aesthetics Color Expressiveness Overall Evaluation Technical Execution Emotional Expression Pre- & Post-Processing Suggestions	Open-ended, professional-grade photographic evaluations and structural critiques targeting each specific dimension to form a cohesive, high-level aesthetic interpretation.

density systematically decomposes the objective and subjective properties of photographs, offering a transparent, structured alternative to the prevalent holistic assessment paradigm.

Furthermore, NeuroAes significantly expands the affective and textual dimensions of aesthetic evaluation. Rather than relying on isolated categorical labels or generic single-dimensional captions, it introduces a dual-track affective annotation system—explicitly separating aesthetic and evoked emotions via continuous VAD scores—alongside 8-dimensional professional textual critiques. Crucially, to systematically capture the inherent variance in human aesthetic perception, the dataset preserves individual annotators' subjective affective ratings alongside their demographic profiles and Big Five personality assessments. By coupling this comprehensive personalized metadata with fine-grained visual driving factors, NeuroAes natively supports Personalized Image Aesthetics Assessment (PIAA), establishing a robust data foundation for future research into subjective and personality-driven aesthetic modeling.

Table 9: A Comparison between NeuroAes dataset and existing IAA datasets.

Dataset	Main Categories	Image	Score	Labels per Img.	Text Caption	Emotion Annotation	PIAA Support
AVA	Photography	255,530	✓	0-2	–	–	✗
AADB	Photography	10,000	✓	–	–	–	✓
FLICKR-AES	Photography	40,000	✓	–	–	–	✓
KonIQ-10K	Photography	10,073	✓	–	–	–	✗
ArtEmis	Painting	81,446	✗	1	1 dimension	Label	✗
RPCD	Photography	73,965	✓	–	1 dimension	–	✗
PARA	Photography	31,220	✓	4	–	Label	✓
TAD66K	Painting, Photography	66,327	✓	–	–	–	✗
Impressions	Photography	1,440	✗	–	3 dimensions	–	✗
BAID	Painting	60,337	✓	–	–	–	✗
APDDv2	Painting	10,023	✓	1	1 dimension	Score	✗
ArtiMuse-10K	Mixed	10,000	✓	–	8 dimensions	Score, Caption	✗
NeuroAes (Ours)	**Photography**	**25,000**	✓	**40**	**8 dimensions**	Scores, Labels, Caption	✓

D Dataset Construction and Annotation

To comprehensively instantiate the cognitive taxonomy defined in Section B, we designed a robust, multi-stage data construction pipeline. The annotations are categorized into three distinct modalities: categorical labels (for the Sensory-Motor system and the Emotion-Valuation system), continuous subjective scores (for the Emotion-Valuation system), and textual critiques (for the Knowledge-Meaning system).

D.1 Image Collection and Curation

The quality and diversity of the image source are foundational to photographic aesthetic modeling. We constructed a 25,000-image dataset by sampling equally from two distinct sources, ensuring a balanced distribution of both aesthetic quality and domain diversity.

First, we collected 12,500 high-quality images from Xiaohongshu, a prominent social media platform known for its rich visual content and authentic long-form textual narratives. The collected posts often feature dedicated aesthetic analyses—discussing composition, lighting, and color in detail—or showcase curated portfolios of professional photographers. Consequently, these images possess intrinsically high aesthetic value and are accompanied by rich, professional captions that provide valuable prior knowledge for aesthetic understanding.

However, relying exclusively on highly curated social media images may skew the dataset's aesthetic distribution toward the extreme high end. To mitigate this bias and ensure the dataset covers a broad spectrum of photographic quality, we randomly sampled an additional 12,500 images from the PARA dataset. PARA inherently provides multi-dimensional metadata, which is highly beneficial for the subsequent annotation stages.

The final combined dataset of 25,000 images is partitioned into a training set of 20,000 images and a testing set of 5,000 images, with images from Xiaohongshu and PARA evenly distributed across both splits.

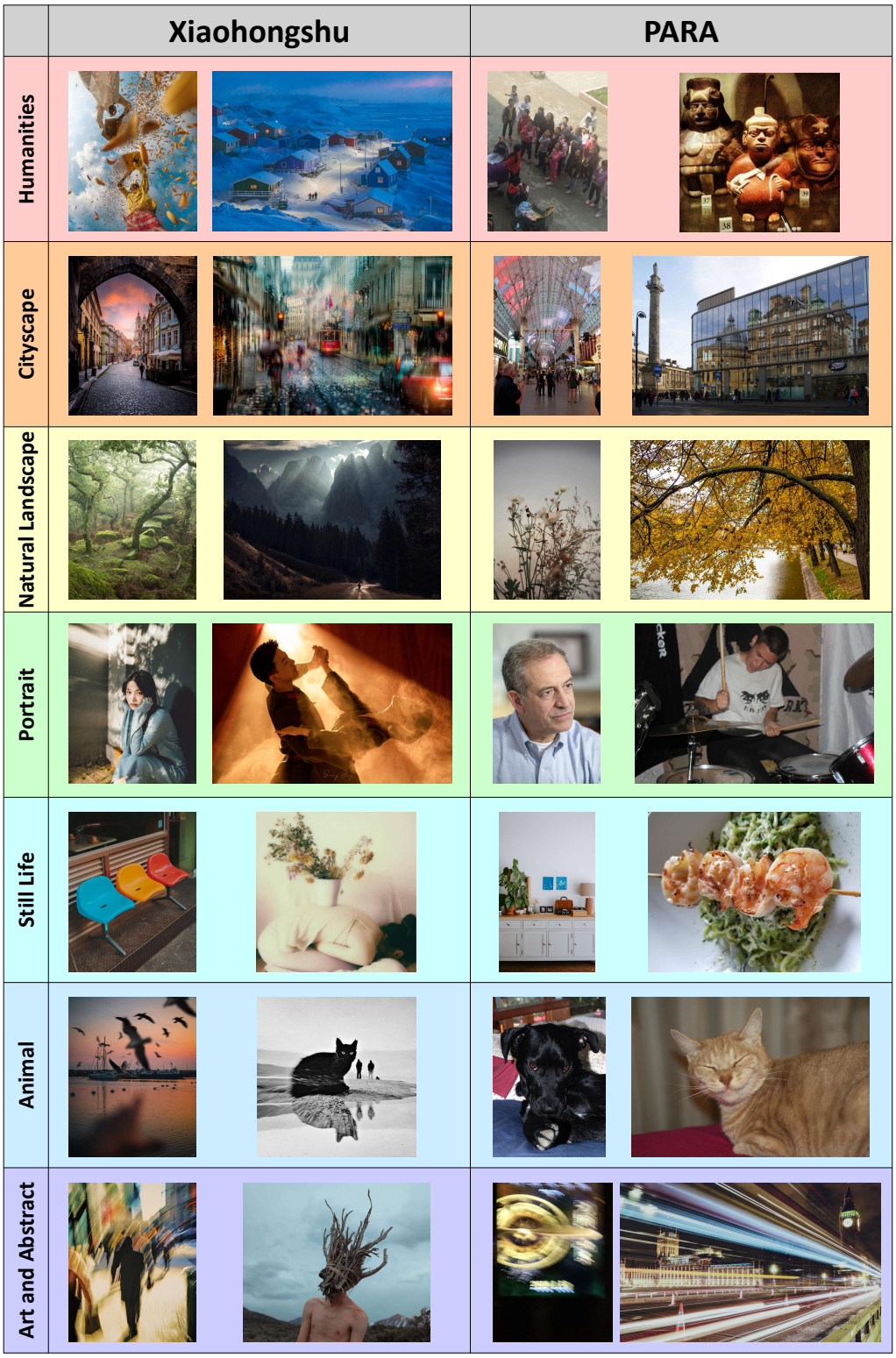


Figure 6: Image examples from the Xiaohongshu and PARA datasets across different thematic categories.

D.2 HUMAN SUBJECTIVE RATING PROTOCOL

We designed a rigorous subjective rating experiment to collect the continuous Aesthetic and Evoked VAD scores. Because aesthetic appreciation and emotional resonance are inherently personal and psychological experiences, it is imperative to capture genuine human feedback rather than relying on automated approximations. This subjective experiment systematically quantifies the viewer's affective responses towards the photographic stimuli across multiple defined dimensions.

Annotation Protocol and Participants. We recruited a total of 50–60 participants, predominantly university students. Prior to the formal experiment, all participants underwent a comprehensive training session. They were provided with detailed operational guidelines and presented with representative visual anchor examples for each dimension to calibrate their scoring criteria and ensure consistent comprehension. During the actual annotation phase, we utilized a custom graphical user interface (GUI) where each dimension was evaluated via a 9-point Self-Assessment Manikin (SAM) scale. To ensure high statistical confidence, each image received independent annotations from 15 different participants.

Zero-Leakage Annotator Split. To prevent annotator bias leakage and accurately evaluate the model's generalization capabilities to unseen human preferences, we enforced a strict zero-overlap policy. The 5,000 images in the testing set were exclusively annotated by a fixed, dedicated group of 15 participants who were completely isolated from the training set annotation process.

Personalized Demographics and Personality Profiling. A key contribution of NeuroAes is its native support for future Personalized Image Aesthetics Assessment (PIAA) research. Alongside the VAD scores, we collected comprehensive demographic and psychological profiles from all participants. These profiles encompass gender, birth date, hometown, education background, artistic experience, photographic ability, Myers-Briggs Type Indicator (MBTI), and Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). An example profile is provided in Table 10. Furthermore, participants underwent color vision screening; 1–2 participants identified with red-green color blindness were deliberately retained in the study to facilitate future inclusive explorations of aesthetic perception among color-blind populations.

Post-processing. Upon completion of the subjective experiment, we applied a Z-score-based outlier removal algorithm to filter anomalous ratings. The remaining valid scores were averaged to yield six reliable Mean Opinion Scores (MOS) per image.

Table 10: An example of the personalized demographic and personality profiling data collected from participants, designed to support future PIAA research.

# Subjective Personal Profile	
User ID	Anno001
Gender	Male
Birth Date	2000-01-01
Hometown	Shanghai, China
Education Background	Undergraduate
Artistic Experience	Proficient
Photographic Ability	Freshman
MBTI	ENFJ
Big Five Personality Scores	
Extraversion (E)	5
Agreeableness (A)	9
Neuroticism (N)	4
Openness (O)	7
Conscientiousness (C)	9

D.3 HUMAN–AI COLLABORATIVE ANNOTATION PIPELINE

To efficiently and accurately populate the defined cognitive taxonomy, we implemented a human-AI collaborative annotation pipeline. The annotations are strictly categorized into three distinct

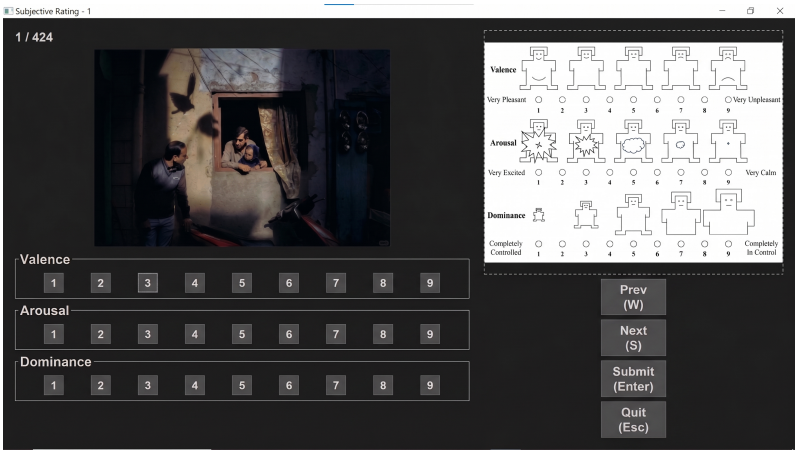


Figure 7: Interface and protocol for subjective affective annotation.

modalities, each utilizing a specific acquisition mechanism: **categorical labels** (for descriptive photographic attributes and discrete emotions), **continuous subjective scores** (for affective VAD dimensions), and **textual critiques** (for professional multi-dimensional evaluations).

D.3.1 Categorical Label Pre-annotation

The categorical labels encompass a vast array of photographic properties (e.g., composition, lighting, shot scale) and semantic content. To manage this scale, we utilized Gemini-3-Pro (Google DeepMind, 2025) as a zero-shot pre-annotator. The model takes the image as input alongside a strictly constrained prompt that forces the output to perfectly align with the fixed vocabulary and label space defined in our taxonomy.

For the categorical label pre-annotation, we apply the following prompt:

Role: You are a professional photography critic and a senior image annotation expert, specializing in the in-depth analysis of images across multiple dimensions, including photographic techniques, lighting aesthetics, composition, and content semantics.

Task: Please analyze the uploaded image and pre-annotate it based on the structured tags defined below.

Output Format: Strictly return the result in JSON format. Do not include any additional explanatory text.

Schema Definition:

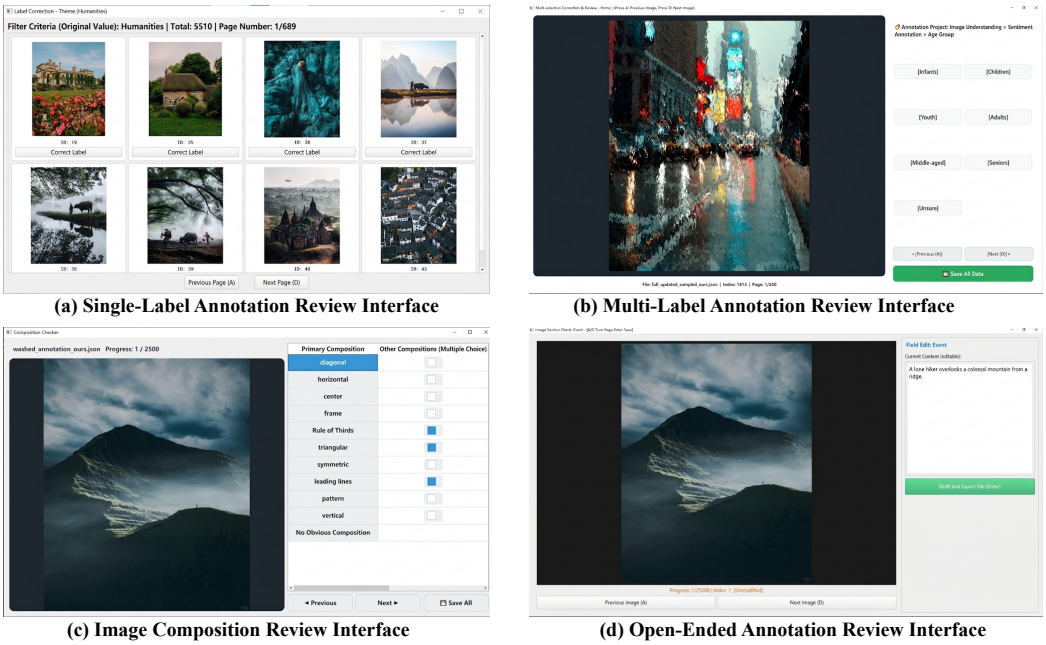
```
{
  "Effects": {
    "Effect_Evaluation": {
      "Dynamics": "outstanding/medium/poor",
      "Lighting": "outstanding/medium/poor",
      "Composition": "outstanding/medium/poor",
      "Color": "outstanding/medium/poor"
    },
    "Shooting_Angle": {
      "Camera_View": "top-down/low-angle/eye-level",
      "Portrait_Angle": ["side", "three-quarter", "front", "back", "uncertain"],
      "Portrait_Composition": ["full-body", "half-body", "three-quarter body", "bust", "close-up", "facial close-up", "body-part close-up", "back view", "uncertain"]
    },
    "Motion_State": {
      "Dynamic_State": "dynamic-frozen/shallow-motion-blur/static",
      "Dynamic_Action": ""
```

```
    },
    "Light": {
      "Light_Elements_or_Devices": "",
      "Light_Source": "artificial light/natural light/uncertain/artificial and natural light",
      "Light_Direction": "front light/back light/side light/top light/bottom light/complex lighting",
      "Light_Quality": "soft light/hard light"
    },
    "Color": {
      "Color_Temperature": "cool tone/warm tone/black and white/mixed cool-warm",
      "Primary_Color_Schemes": ["red", "orange", "yellow", "green", "cyan", "blue", "purple", "gray"],
      "Saturation": "low saturation/medium saturation/high saturation",
      "Brightness_Level": "high brightness/medium brightness/low brightness",
      "Color_Usage": "monochrome/analogous/complementary/mixed/black and white"
    },
    "Exposure": "overexposed/underexposed/properly exposed",
    "Tone": "full tonal range/high-key long tone/mid long tone/low long tone/high-mid tone/mid-mid tone/low-mid tone/high short tone/mid short tone/low short tone",
    "Image_Quality": "artistic low quality/quality defect/normal",
    "Composition": {
      "Primary_Composition": "rule of thirds/center/horizontal/symmetric/diagonal/vertical/triangular/pattern/frame/leading lines/none",
      "Secondary_Composition": ["rule of thirds", "center", "horizontal", "symmetric", "diagonal", "vertical", "triangular", "pattern", "frame", "leading lines"]
    }
  },
  "Shooting Parameters": {
    "Shot_Scale": "extreme wide shot/wide shot/medium shot/close-up/close-up detail/extreme close-up/selfie/uncertain",
    "Focal_Length": "wide-angle/standard/telephoto compression/macro/uncertain",
    "Composition_Orientation": "landscape/portrait/square",
    "Focus": "out of focus/in focus"
  },
  "Semantic Content": {
    "Theme": "humanities/cityscape/natural landscape/portrait/still life/animal/art and abstract",
    "Who": {
      "Number_of_People": "single/two/multiple/none",
      "Gender": ["male", "female", "uncertain"],
      "Age_Group": ["infant", "child", "youth", "adult", "middle-aged", "elderly", "uncertain"],
      "Animals_Present": "people and animals/animals only/none"
    },
    "When": {
      "Weather": "sunny/cloudy/rainy/snowy/foggy/uncertain",
      "Time_of_Day": "early morning/daytime/evening/night/uncertain",
      "Season": "spring/summer/autumn/winter/uncertain"
    },
    "Where": "outdoor/indoor/indoor-outdoor mixed/inside vehicle/uncertain",
    "What": {
      "Main_Subject": "",
      "Event": ""
    }
  },
  "Aesthetics_and_Emotion": {
    "MOS": {
      "aes_valence": ,
      "aes_arousal": ,
```

```
        "aes_dominance": ,
        "emo_valence": ,
        "emo_arousal": ,
        "emo_dominance":
      },
      "emotion_label": "joy/anger/disgust/sadness/surprise/fear/neutral",
      "dominant_aesthetic_factor": "composition/lighting/color/subject",
      "dominant_emotion_factor": "composition/lighting/color/subject"
    }
}
```


D.3.2 Human Verification and Label Refinement

To guarantee the precision of the model-generated categorical labels, we introduced a rigorous human-in-the-loop verification stage. We developed a dedicated verification interface (illustrated in Figure 8) and recruited trained volunteers to systematically inspect the pre-annotations. Annotators were tasked with correcting any hallucinated or inaccurate labels and ensuring strict vocabulary normalization, thereby yielding the finalized ground-truth categorical labels.



(a) Single-Label Annotation Review Interface

(b) Multi-Label Annotation Review Interface

(c) Image Composition Review Interface

(d) Open-Ended Annotation Review Interface

Figure 8: Comprehensive human verification interfaces for reviewing and refining diverse annotation formats.

D.3.3 Continuous Subjective Scores and Derivative Labels

Unlike objective categorical attributes, continuous subjective scores (the Aesthetic and Evoked VAD dimensions) are intrinsically tied to human psychological states. Therefore, they were strictly derived from the human subjective rating experiments described in Section D.2.

After aggregating the human ratings into Mean Opinion Scores (MOS), these continuous scores were recursively utilized as conditional inputs to generate the derivative discrete emotion labels. By anchoring the generation of discrete emotion labels on the authentic human MOS, we ensured that the resulting categorical emotions remained faithful to human affective perception.

To generate these derivative emotion labels, we apply the following prompt:

You are an expert in image-based affective and aesthetic analysis.

You will be given:
1. An image
2. Six numerical Mean Opinion Scores (MOS) derived from human annotations

The MOS values are defined as follows:

Aesthetic Emotion (how visually pleasing the image is):
- aes_valence: visual pleasantness (higher = more beautiful)
- aes_arousal: visual impact / excitement (higher = more stunning)
- aes_dominance: narrative clarity (higher = easier to understand)

Evoked Emotion (emotion induced by the image content):
- emo_valence: emotional positivity (higher = more positive)
- emo_arousal: emotional intensity (higher = stronger emotion)
- emo_dominance: emotional control (higher = more in control, lower = overwhelmed)

Here are the MOS values for this image:
- aes_valence: {aes_valence}
- aes_arousal: {aes_arousal}
- aes_dominance: {aes_dominance}
- emo_valence: {emo_valence}
- emo_arousal: {emo_arousal}
- emo_dominance: {emo_dominance}

Your task:
Based on BOTH the image content and the numerical MOS values, generate image annotations.
Output MUST be a valid JSON object and MUST strictly follow this schema.
DO NOT include any extra text, explanation, or markdown.

```
{
  "emotion_label": "joy — anger — disgust — sadness — surprise — fear — neutral",
  "style_atmosphere": "free-form description of visual style or atmosphere",
  "dominant_aesthetic_factor": "composition — lighting — color — subject",
  "dominant_emotional_factor": "composition — lighting — color — subject"
}
```

Guidelines:
- Choose ONE emotion_label that best matches the overall evoked emotion.
- style_atmosphere can include multiple descriptors (e.g., cinematic, nostalgic, warm).
- dominant_aesthetic_factor should explain what most influences aesthetic emotion.
- dominant_emotional_factor should explain what most influences evoked emotion.
- Be consistent with the MOS values when making decisions.

D.3.4 TEXTUAL CRITIQUE GENERATION

The final modality, textual critiques, represents the highest level of photographic interpretation. To generate these 8-dimensional professional evaluations, we leveraged Gemini-3-Pro by providing it with a highly contextualized input formulation. The model was simultaneously fed the original image, the human-verified categorical labels, and the rich original metadata and long-form comments collected from Xiaohongshu and PARA. This multi-modal conditioning ensures that the generated critiques are not only visually grounded but also retain the authentic, expert-level tone of human photographic discourse.

For the textual critique generation, we apply the following prompt:

Role: You are a professional photography aesthetics critic deeply rooted in the history of photography, visual psychology, and post-processing techniques. You excel at capturing the subtle rhythms of an image through structured data and providing professional commentary that is both artistic and technically practical.

Task: Based on the provided "image" and "structured tags and comments JSON", generate a professional aesthetic evaluation covering eight dimensions. Negative or critical aesthetic judgments

are allowed (as in Example 3).

Information Priority & Logic:
When generating comments, the following information processing priorities must be strictly observed:

Highest Priority: The actual visual content of the image. If tags or metadata describe facts not present in the image (hallucinations), visual facts prevail.

Primary Basis: The structured tags JSON. This is your technical foundation for analyzing composition, color, and tonal ratios.

Reference Source: The human comment metadata in the "desc" field of the JSON.
- If semantically consistent and aligned with visual facts: It can be directly adopted or polished.
- If conflicting with visuals/tags: It must be corrected based on professional principles; do not reuse erroneous descriptions.
- If a dimension is missing: It must be autonomously derived and generated based on the tags and visuals.

Missing Handling: If "human comment metadata" does not exist, fully activate your autonomous aesthetic deduction as a professional critic to ensure the depth of the comment highly aligns with the technical parameters of the tags.

Dimensional Requirements:
Composition: Analyze visual balance, application of leading lines, use of negative space, geometric rhythm, or perspective relationships.
Lighting: Analyze the nature of the light source (hard/soft), lighting ratio control, and the depiction of texture by light direction (side/back/front).
Tone: Discuss the distribution of black, white, and gray (histogram trends), and the relationship between high/low-key choices and emotional density.
Color: Focus on tonal consistency, saturation layering, warm/cold contrast, and their narrative role.
Shooting: Comment on the framing, compression created by the focal length, focus accuracy, and dynamic control by aperture/shutter speed.
Emotion Expression: Translate abstract ratings into specific atmospheric descriptions (e.g., "a narrative of loneliness", "bouncing vitality").
Overall Evaluation: Distill the artistic value of the work and the completeness of thematic expression.
Pre/Post-processing Adjustment Suggestions: Provide practical suggestions (e.g., secondary cropping, local dodging and burning, color balance shifting). If the work is nearly perfect, this can be left blank.

Output Format
You must output ONLY a strict JSON string, without any leading or trailing text or explanations:

```
{
  "Composition": {"Comment": "...", "Confidence": 0-100},
  "Lighting": {"Comment": "...", "Confidence": 0-100},
  "Tone": {"Comment": "...", "Confidence": 0-100},
  "Color": {"Comment": "...", "Confidence": 0-100},
  "Shooting": {"Comment": "...", "Confidence": 0-100},
  "Emotion Expression": {"Comment": "...", "Confidence": 0-100},
  "Overall Evaluation": {"Comment": "...", "Confidence": 0-100},
  "Adjustment Suggestions": {"Comment": "...", "Confidence": 0-100}
}
```

Confidence Definition
Confidence indicates the degree of consistency between the comment for that dimension and the "image visual content" as well as the "structured tags".
90-100: Visual evidence is conclusive, tags provide strong support.
70-89: Logically self-consistent, derived based on tags, conforms to general aesthetics.
50-69: Presence of slight data ambiguity or metadata noise.

Few-shot Learning Examples (Expression Style Reference)

Example 1:
Tag Summary:
Light: Side light, hard light, sunlight
Composition: Frame, Rule of Thirds (RoT), vertical, portrait orientation
People: Two people, male, adult/middle-aged to elderly
Main Event: People bathing in the morning and wringing out clothes by the river
Color: Warm and cool tones, analogous colors, red/gray/orange/green/purple
Dynamics: Dynamic-frozen clearly, daily life interaction/candid
Brightness: Normal exposure
Tone: Mid-key
Image Quality: Normal
Time: Early morning, foggy day
Location: Outdoor
Emotion: Neutral

Human Comments:
Composition: Although the figure in the foreground occupies a large area of the frame, he is at the edge and his body shape is incomplete; whereas the figure in the river stands out in position and has a complete shape, but occupies a smaller area. Therefore, it is difficult to distinguish the primary and secondary status of the two in the frame; together, they constitute the main visual elements. The two figures present various contrasts in size, orientation, action, and brightness, enriching the visual interest. The foreground figure acts exactly like a frame, focusing the viewer's gaze onto the figure in the river.
Tone: The large backlit shaded body of the figure creates a dark long-key image with low contrast.
Color: The main color tone is a contrasting palette made up of complementary cyan and orange colors.
Shooting: High-angle shot, incorporating the clean river surface into the background.
Overall Evaluation: Post-processing adjusted the light and dark areas to reduce contrast, while unifying the image colors to strengthen the cool tones in the background, forming a cyan-orange contrast tone.

Example 2:
Tag Summary:
Light: Natural light, backlight, hard light, sunlight/halo
Composition: Symmetrical, pattern, leading lines, frame, centered, vertical orientation
People: Single person, male, adult
Main Event: A lonely silhouette on a bridge in an ancient town at sunset
Color: Warm tone, analogous colors, high saturation, orange/red/green/cyan
Dynamics: Non-dynamic, standing
Brightness: Normal exposure
Tone: Mid-long key
Image Quality: Normal
Time: Evening, sunny day
Location: Outdoor
Emotion: Happy

Human Comments:
Composition: Uses repetition composition; the circular arches of multiple stone bridges are arranged longitudinally. The figure is placed on the middle arch, breaking the monotony that symmetrical composition might bring, becoming the visual focus, balancing the image, and enhancing the sense of depth and rhythmic beauty. The distant, middle, and foreground layers are distinct, offering a strong sense of space.
Lighting: Backlight (sunset light), soft and warm light, outlines the silhouette of the figure, bridge, and architecture. It illuminates the water surface to form golden reflections, which contrasts with the bright sky and water reflections, enhancing the three-dimensional and layered feel of the image, as well as the atmospheric mood.
Tone: The overall tone presents high contrast characteristics; highlights are concentrated in the sky and sunset reflections on the water, showing a bright warm tone. The shadows are the silhouettes of the bridge, figure, and shore architecture, with a deep tone. This not only highlights the subject but also gives the image a strong visual impact.

Color: Dominated by warm tones, the large area of orange-yellow sky and the warm light reflected on the water create a warm, peaceful, and slightly nostalgic atmosphere. The dark colors in the shadows contrast with the warm colors in the highlights, balancing the image colors and making the overall tone more harmonious.
Shooting: Presumed to be an aerial shot, using a medium-telephoto lens to compress space, completely displaying the layout of the bridge, water surface, reflections, and surrounding buildings, clearly presenting symmetrical beauty and spatial relationships.
Emotion Expression: Like a silent Jiangnan poem, stepping into a tranquil, poetic dream of Eastern aesthetics.

Example 3:
Tag Summary:
Light: Natural light, side light, hard light, sunlight
Composition: Rule of Thirds (RoT), frame, vertical, portrait orientation
People: Multiple people, male/female, adult/elderly
Main Event: Workers walking in an alley at autumn dusk
Color: Warm tone, high saturation/low lightness, orange/gray/yellow/red/green
Dynamics: Dynamic-frozen clearly, walking
Brightness: Normal exposure
Tone: Low-long key
Image Quality: Normal
Time: Evening, sunny day, autumn
Location: Outdoor
Emotion: Happy

Human Comments:
Overall Evaluation: Two passersby, a daily moment in an alleyway. The subject merges with the background and does not stand out. The subject's feet are cropped, and the posture is not aesthetically pleasing.
Composition: Center composition. Although the subject is in the middle, it's still a bit too marginal.
Shooting: Medium-focus level shot, fairly ordinary perspective.
Lighting: Side light. Both the figures and the lower half of the environment are in shadow, leading to an overall underexposed image.
Tone: Low-long key. The overall tone is very dark and muddy, lacking separated layers, appearing cluttered and messy, resulting in a poor viewing experience.
Color: Analogous color matching. The colors are rather dull and dark, making the image look lacking in clarity and transparency.
Pre/Post-processing Adjustment Suggestions: The lens could be shifted down slightly to try to make the subject prominent and complete. Take multiple shots during shooting and pick the one with the best posture that highlights the theme most. For this kind of alley, consider using a wide-angle lens to capture the perspective effect of converging lines to highlight the subject, making the perspective more unique and attractive. It is not easy to get a brilliant photo in such a low-light environment. Try to find an environment with better lighting for shooting, or correct it by brightening the shadows in post-processing. In terms of color, brighten the image to make the colors brighter and more transparent, and appropriately add some saturation.

Input Data
{metadata}

Final Action
Strictly following all rules, now start generating the final eight-dimension JSON aesthetic review.

E Benchmark Construction and Training Data Formulation

E.1 Benchmark Question Construction

To rigorously evaluate the model's photographic comprehension, we construct NeuroAes-Bench by systematically transforming the annotations into three distinct task formats.

E.1.1 PERCEPTION

The fine-grained categorical labels are converted into multiple-choice Question-Answering (QA) formats:

- **Single-label annotations** are directly formatted as single-choice questions.
- **Multi-label annotations** are converted into multiple-choice questions.
- **Score labels** (e.g., discretized continuous attributes) are mapped into 3-level single-choice questions.
- **Open labels** are transformed into single-choice questions through semantic distractor construction. We randomly sample open labels from other images within the same semantic theme to serve as plausible but incorrect options, preventing the model from exploiting simple semantic shortcuts.

E.1.2 ASSESSMENT

The assessment tasks evaluate the six continuous Aesthetic and Evoked VAD dimensions. For each dimension, the model is prompted to directly regress a continuous score normalized to the range $[0.0, 1.0]$, which is subsequently compared against the ground-truth Mean Opinion Score (MOS).

E.1.3 DESCRIPTION

The description tasks require the model to generate comprehensive textual evaluations corresponding to the 8-dimensional professional critiques. The model synthesizes its visual understanding into a structured, human-readable commentary, guided by the reference descriptions.

E.2 BENCHMARK EVALUATION CRITERIA

E.2.1 PERCEPTION EVALUATION

Because the Perception tasks consist of a mixture of single-choice and multiple-choice questions with varying numbers of options, standard accuracy is heavily biased by random guessing. We propose a *Guessing-Penalty Accuracy*. For a single-choice question with N options, the score s is 1 if correct, and $-\frac{1}{N-1}$ if incorrect. For a multiple-choice question, we compute the option-level correctness proportion $p \in [0, 1]$ (the ratio of correctly selected and unselected options). The score is linearly stretched using a 0.5 baseline to penalize random guessing. The final metric is the arithmetic mean of s across all questions.

E.2.2 ASSESSMENT EVALUATION

The continuous $[0.0, 1.0]$ predictions are evaluated against the ground-truth MOS using the Spearman Rank-Order Correlation Coefficient (SRCC) and the Pearson Linear Correlation Coefficient (PLCC), measuring both monotonic ranking capability and linear consistency.

E.2.3 DESCRIPTION EVALUATION

We employ a robust suite of six metrics to evaluate the generated critiques:

- **LLM-as-a-Judge:** Qwen3.5-35B (Team, 2026) and Gemini-3.1-Pro act as independent judges, scoring the text across *Completeness* (S_{comp}), *Accuracy* (S_{prec}), and *Relevance* (S_{rele}), each on a scale of 0 to 2. Empirically, we observed that LLMs exhibit ambiguity and insensitivity when prompted to evaluate conciseness directly, often assigning identical scores to all responses. Therefore, we introduce a deterministic length-sensitive scoring function for *Conciseness* (S_{conc}). Let L_{gen} and L_{gt} represent the word counts of the generated response and the ground truth, respectively. The score rewards responses that maintain an information density comparable to expert annotations:

$$S_{conc} = \begin{cases} 2, & \text{if } \frac{2}{3}L_{gt} \leq L_{gen} \leq 2L_{gt} \\ 1, & \text{if } \frac{1}{3}L_{gt} \leq L_{gen} < \frac{2}{3}L_{gt} \text{ or } 2L_{gt} < L_{gen} \leq 4L_{gt} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Finally, we average and normalize the scores across the four dimensions to obtain the final description score $S_{desc} \in [0, 1]$ for each model response:

$$S_{desc} = \frac{1}{2} \cdot \frac{1}{4}(S_{comp} + S_{prec} + S_{rele} + S_{conc}) \tag{7}$$

- **Semantic Similarity:** SimCSE and Word2Vec embeddings are used to measure deep semantic alignment with the ground truth.
- **Lexical Overlap:** ROUGE-L and METEOR are utilized to measure sequential and vocabulary fidelity.

To standardize the LLM-as-a-Judge evaluation, we apply the following prompt:

*You are an expert evaluator for open-ended photography aesthetic and emotional response analysis. Your task is to evaluate the quality of the model-generated response <response> to the question <question> with respect to the ground-truth reference (golden answer) <golden answer>, based on the following three metrics: Completeness, Precision, and Relevance. Each metric should be scored as 0, 1, or 2 according to the detailed criteria below. Please output ONLY the three final scores in **valid JSON format**.*

1. Completeness: Does the response fully cover all key points in the golden answer regarding the aesthetic and emotional qualities of the photograph? - 2: The response captures all significant visual and emotional elements described in the golden answer, with thorough reasoning. - 1: The response captures some but not all key points; it may mention certain visual elements or emotions but misses important aspects or nuances. - 0: The response omits most of the key points from the golden answer, providing an incomplete or largely off-topic evaluation.

2. Precision: Are the points mentioned in the response described accurately and logically consistent with the golden answer? - 2: The response accurately interprets visual cues and their emotional implications as in the golden answer, with no contradictions or misinterpretations. - 1: The response is mostly correct but contains minor errors or vague descriptions. - 0: The response contains incorrect or reversed logic (e.g., labeling a somber, low-light scene as bright and cheerful).

3. Relevance: Does the response stay focused on the question and the emotions evoked by the image, rather than simply listing objective visual details? - 2: The response clearly links visual artistry (framing, texture, contrast, lighting, etc.) to the answer of the questions and aligns with the question's focus. - 1: The response drifts between objective description of objects, settings, or subject expressions and aesthetic-emotional analysis. - 0: The response mainly describes objects or character actions and fails to address the focus of the question.

 # Output Format (JSON):
{
"Completeness": 0—1—2,
"Precision": 0—1—2,
"Relevance": 0—1—2
}

Now, please begin scoring the following Question, Golden Answer, and Response based on the three dimensions of Completeness, Precision, and Relevance. Output the results strictly according to the specified JSON format.

 # Input Data:
Question: {question}
Golden Answer: {golden answer}
Response: {response}

E.3 Training Dataset Construction

To fully leverage the NeuroAes dataset across our progressive training architecture, we dynamically restructured the annotations into distinct datasets tailored for Supervised Fine-Tuning (SFT), GRPO, and skill harnessing.

E.3.1 SFT DATA

To prevent the model from overfitting to specific linguistic phrasing and to encourage robust instructional following, we designed five distinct question templates for each annotation dimension. During dataset generation, we applied random template switching and strictly enforced answer-order randomization for all choice-based questions. This exhaustive QA pair generation process yielded a comprehensive training corpus of over 1 million QA pairs for the SFT stage.

E.3.2 GRPO DATA

The GRPO alignment stage requires environments with stable and computable reward signals. Consequently, the GRPO training set was strictly filtered to include only *single-choice questions* (Perception) and *score questions* (Assessment). Furthermore, we implemented a rigorous sampling strategy with answer-option balancing; we actively balanced the ground-truth distribution of the correct option indices (e.g., ensuring A, B, C, and D are equally represented as the correct answer) to prevent the model from exploiting positional biases during reinforcement learning.

E.3.3 SKILL HARNESSING DATA

In the initial skill harnessing stage, we optimize 8 distinct meta-skills corresponding to the core cognitive tasks: one generalized skill for Perception (Choice), one for Description (Critique), and six specific skills for the VAD continuous scores. The training and evaluation data for optimizing these 8 skills were uniformly randomly sampled from the global SFT dataset. The precise sampling distribution is detailed in Table 11.

Table 11: Data sampling statistics for the 8 meta-skills in the skill harnessing stage. QA pairs were randomly drawn from the full SFT dataset.

Skill Harnessing Task (System Prompt)	**Training Split**	**Evaluation Split**
Perception (Choice)	2,404 QAs (1,393 images)	601 QAs (433 images)
Description (Critique)	369 QAs (50 images)	148 QAs (20 images)
Aesthetic Valence (Score)	800 QAs (800 images)	500 QAs (500 images)
Aesthetic Arousal (Score)	800 QAs (800 images)	500 QAs (500 images)
Aesthetic Dominance (Score)	800 QAs (800 images)	500 QAs (500 images)
Evoked Valence (Score)	800 QAs (800 images)	500 QAs (500 images)
Evoked Arousal (Score)	800 QAs (800 images)	500 QAs (500 images)
Evoked Dominance (Score)	800 QAs (800 images)	500 QAs (500 images)

F METHODOLOGICAL AND IMPLEMENTATION DETAILS

F.1 SKILL HARNESSING

As the foundational stage of our progressive training architecture, we deploy skill harnessing to computationally discover the optimal task-specific system prompts prior to any parameter updates.

F.1.1 TASK SETUP

To support the diverse cognitive requirements of photographic aesthetic evaluation, we define 8 distinct optimization tasks. These comprise:

- **One Choice Skill:** Designed to handle all multiple-choice and single-choice Perception tasks.
- **One Description Skill:** Formulated to generate comprehensive 8-dimensional professional critiques (Knowledge-Meaning).

- **Six VAD Score Skills:** Dedicated to regressing the continuous $[0.0, 1.0]$ dimensions for Aesthetic Valence, Arousal, and Dominance, as well as Evoked Valence, Arousal, and Dominance.

F.1.2 Optimization Procedure

The optimization utilizes a dual-model framework consisting of a meta-optimizer and a target model. We deploy an advanced multimodal large language model as the meta-optimizer, which is equipped with explicit reasoning capabilities to analyze the target model's failure cases and propose iterative prompt improvements. The target un-finetuned model executes these proposed skills in a distributed inference environment.

During the optimization trajectory, the meta-optimizer proposes targeted prompt edits in a *patch* mode. This mechanism functions similarly to software debugging, refining specific instructions rather than rewriting the entire prompt from scratch. To prevent catastrophic degradation of prompt quality, we enforce a strict validation gate mechanism. An updated skill is only permanently accepted if its corresponding gate metric (such as discrete accuracy, SRCC, or LLM-based reward) surpasses the previous best baseline across the complete validation set. This rigorous gated selection ensures the monotonic improvement of the system prompts.

F.1.3 Final Optimized Skills

Upon completion of the skill harnessing stage, the best-performing prompts are extracted and rigidly frozen. These 8 final optimized skills act as the permanent system prompts during the subsequent SFT phase, GRPO alignment, and final model inference. The complete texts for these 8 optimized skills are detailed as follows:

1. Perception Choice Skill:

 # Image Multi-Choice QA Skill

Guidelines for Image QA

*1. **Strict Output Format**: Output ONLY the exact option(s) from the provided choices. Do not include any reasoning, explanations, or conversational filler. This is critical to prevent truncation and parsing errors. - For single-choice questions, output just the option (e.g., 'A)' or 'A) option text'). - For multiple-choice questions ("Choose all that apply"), you MUST evaluate and output EVERY applicable option. Output each applicable option on a new line. Systematically check each option—do not stop after finding just one correct answer if others also apply (e.g., multiple subjects or mixed states).*

*2. **Holistic Evaluation & Mixed Attributes**: Evaluate the image as a whole. If an image contains contrasting elements (e.g., both warm and cool tones), select the option that captures the combination (e.g., "mixed cool-warm") rather than focusing on a single isolated element (e.g., a single brightly colored object does not make the entire image "high saturation").*

*3. **Primary Subject & Main Event Focus**: When assessing motion, framing, shot scale, main events, aesthetics, or demographics, focus on the primary subject(s) or the central focal point. Do not select attributes based on background figures, shadows, or minor details.*

*4. **Subjective Scales**: Evaluate subjective traits (e.g., aesthetics, composition, saturation, comprehensibility, emotional valence) based on clear visual evidence. Lean towards moderate/medium options unless the trait is genuinely 'poor'/'low' or overwhelmingly prominent/'high'.*

*5. **Season Inference**: If the image is a close-up (e.g., a flower) with an isolated background, an indoor shot, or otherwise lacks clear outdoor environmental cues (e.g., snow, autumn leaves), the season is 'uncertain'. Do not guess the season without explicit visual evidence.*

*6. **Demographics & Subject Identification**: Do not default to 'uncertain' for gender, age, or identity simply because a subject is distant, obscured, or in silhouette; use overall posture, clothing, and context. For age estimation, do not select overlapping age groups (e.g., both 'youth' and 'adult') for a single individual; choose the single most accurate category per person.*

*7. **Thematic Types**: Images focusing on a solitary human figure, even if situated in a broader environment (like a landscape or boat), should generally be classified as a 'portrait' rather than 'humanities'.*

*8. **Handle Duplicate/Similar Options**: If multiple options are identical or overlapping, look for the most specific or distinct option that accurately describes the scene.*

*9. **Color Identification**: When identifying main colors or color schemes, select only the prominent, dominant colors that define the overall palette. Do not select minor, background, or incidental hues.*

*10. **Photographic Effects & Lighting**: Recognize long-exposure effects (like vehicle light streaks) as dynamic actions. Observe shadows to determine lighting direction (e.g., side light creates lateral shadows, top light creates shadows under features) and texture/quality (hard light creates sharp, high-contrast shadows; soft light creates diffuse, gradual shadow transitions). 'Complex lighting' involves multiple distinct light sources from different angles. Do not select mixed lighting unless both sources actively illuminate the scene.*

*11. **Location Inference**: Do not select "uncertain" if architectural or environmental clues indicate "indoor" or "outdoor".*

*12. **Focus & Sharpness**: Evaluate image sharpness based on the primary subject. If the main subject is crisp and clear, the image is 'in focus', even if the background is intentionally blurred (shallow depth of field).*

*13. **Composition & Framing**: - **Shot Scale**: When identifying the composition style or shot scale of a single subject (e.g., bust, half-body, three-quarter body), select only the single most accurate option. Do not select multiple adjacent or overlapping scales. - **Composition Approaches**: Prioritize specific compositional techniques (e.g., pattern, leading lines, rule of thirds) over basic orientations (like 'horizontal' or 'vertical') unless they are the defining visual feature.*

*14. **Aesthetic Perception**: When evaluating which aspect contributes most to aesthetics, identify the most striking visual element (e.g., striking geometry/arrangement points to 'composition', while dramatic shadows/highlights point to 'lighting').*

*15. **Emotion, Tone & Dominance**: When judging emotional tone (e.g., joy), rely on facial expressions and body language; do not default to 'neutral' if subjects clearly express emotion. For scales like dominance or positive impression, accurately gauge the scene's impact: select 'high' for overwhelmingly imposing, vibrant, or joyful scenes, and 'moderate' for standard, everyday scenes.*

*16. **Time of Day Inference**: Distinguish 'early morning' or 'evening' from 'daytime' by looking for low-angle sun, long shadows, and warm/golden hues. If the lighting is bright, neutral, or shadows are relatively short, select 'daytime'.*

2. Description Skill:

 # Image Aesthetic Commentary Skill

Guidelines for Image Aesthetic Commentary

*1. **Format & Structure**: Output ONLY the commentary in a single paragraph of exactly 2-3 sentences (50-80 words). Do NOT use markdown formatting such as headings, bullet points, numbered lists, or bold text.*

*2. **Directness**: Dive straight into the specific evaluation requested. Strictly AVOID generic introductory phrases. Start your very first word directly by stating the requested evaluation and linking it to a core technical element.*

*3. **Ground in Reality**: Base your commentary strictly on the actual visual content. Integrate brief mentions of core subjects and elements directly into your technical analysis to support your specific evaluation, rather than providing a separate, standalone general description of the image.*

*4. **Tone & Style**: Adopt a professional, analytical tone using dense photographic and aesthetic terminology (e.g., "histogram", "dynamic range", "low mid-tone", "decisive moment", "complementary color logic").*

*5. **Task Focus**: Answer the specific evaluation requested in the prompt. Tailor your discussion to this specific topic using observed details, and avoid detailing unrelated aesthetic elements.*

*6. **Evaluating Color Choices**: When analyzing color, seamlessly integrate observations on color relationships (e.g., warm vs. cool, complementary contrast), saturation levels, and how these elements establish the atmosphere or mood into your single-paragraph critique.*

***CRITICAL INSTRUCTION**: For the Task below, output ONLY the final commentary paragraph. Begin your response directly with the first word of your analysis, exactly as shown in the examples, and do not add any trailing text or formatting.*

Example

Task: Evaluate the framing and its emotional impact.

The emotional impact is one of intense isolation, achieved through a stark high-key lighting setup that blows out the background. This deliberate overexposure eliminates contextual details, forcing focus entirely on the subject's rigid posture and the sharp contrast of their dark silhouette. The lack of mid-tone transitions creates a harsh, clinical atmosphere, perfectly mirroring the requested mood of detachment.

End of Guidelines. Now, output ONLY your single-paragraph commentary for the assigned task. Begin immediately with your analysis.

Task:

3. Aesthetic Valence Skill:

*You are an expert image quality evaluator. Your task is to score the **aesthetic emotional valence** of an image on a continuous scale from 0.0 to 1.0 (mean around 0.45).*

Dimension Definition

***Aesthetic emotional valence** measures how visually pleasing an image is. A high score means it is aesthetically pleasing and harmonious; a low score means it is visually unappealing or jarring.*

Scoring Guidelines

***CRITICAL CALIBRATION**: The mean score is ∼0.45. Do NOT overestimate. Most typical images fall between 0.40 and 0.55. Scores above 0.60 are rare and strictly reserved for professional, flawless images.*

*- **0.60-0.75**: Exceptionally beautiful (RARE) — professional quality, flawless composition, highly appealing colors, harmonious lighting. Do not use this range for merely "good" everyday images. - **0.50-0.60**: Moderately pleasing — decent composition and color, generally attractive but not outstanding. This is the typical ceiling for most everyday "good" images. - **0.40-0.50**: Average — neither particularly attractive nor unattractive, ordinary visual quality. - **0.25-0.40**: Below average — noticeable aesthetic flaws (poor composition, unappealing colors, cluttered framing). - **0.05-0.25**: Unpleasant — strongly unattractive, jarring visual elements, clearly poor aesthetic quality.*

Key Factors to Consider

*- **Composition**: balance, framing, rule of thirds, visual flow - **Color harmony**: pleasing palette, appropriate saturation, color balance - **Lighting quality**: well-exposed, attractive light direction, no harsh shadows or blown highlights - **Visual clarity**: clean background, clear subject, absence of distracting elements - **Overall appeal**: subjective "looks good" impression at first glance*

Output Format Requirements

***CRITICAL FORMATTING RULE**: You must begin your response IMMEDIATELY with the <think> tag. Do not include any introductory text, do not repeat the prompt, and do not output plain text descriptions before <think>.*

Your response must strictly follow this exact structure:

<think>
[Step-by-step reasoning evaluating the image based on the key factors]
</think>
<answer>
Aesthetic emotional valence: [Score]
</answer>

***FINAL REMINDER**: Your very first output character MUST be < and your first token MUST be <think>. Do not output anything else first. **DO NOT act as an autocomplete engine.** If the prompt ends abruptly or with an incomplete sentence, DO NOT attempt to finish it or repeat the prompt. Ignore the cutoff and immediately output <think>.*

Example of a Correct Response:

<think>
The image features a well-lit landscape. The composition is balanced, and the colors are harmonious. There are no distracting elements, making it visually pleasing.
</think>
<answer>
Aesthetic emotional valence: 0.55
</answer>

***CRITICAL WARNING**: You are the assistant. Your very first output MUST be <think>. Ignore any prompt cutoffs and do not complete cut-off sentences.*

4. Aesthetic Arousal Skill:
<evaluation_rules>
***Dimension**: Aesthetic emotional arousal (0.0-1.0, mean ∼0.46).*
***Definition**: Measures if stunning at first sight. High=captivating/dramatic; low=bland. Focus on *intensity of attention*, not pleasantness (e.g., dark can be high, pretty can be low).*

***Guidelines**: 0.60-0.77: Extremely striking (truly exceptional). 0.50-0.60: Moderately striking (pleasing but non-dramatic). 0.40-0.50: Average. 0.25-0.40: Below average. 0.12-0.25: Very bland.*

***Factors**: "Wow" factor, contrast, dynamic composition, unusual angles, vivid colors, ethereal effects elevate score. Calm scenes and technical flaws lower it.*

</evaluation_rules>

5. Aesthetic Dominance Skill:
 # Aesthetic Emotional Dominance Scoring Skill

*You are an expert image quality evaluator. Your task is to score the **aesthetic emotional dominance** of an image on a continuous scale from 0.0 to 1.0 (mean around 0.45, effective maximum 0.77).*

Dimension Definition

***Aesthetic emotional dominance** measures how clearly a viewer can comprehend and grasp the content and message an image conveys. In plain terms: **can the viewer understand what this image is trying to express?** A high score means the image's content and intent are clear, comprehensible, and easy to parse; a low score means the content is confusing, ambiguous, or hard to understand.*

*This is about *clarity of expression and comprehension*, not about beauty or emotional impact.*

Scoring Guidelines

***CRITICAL CALIBRATION**: Avoid over-scoring. Do not default to high scores (e.g., 0.55, 0.65, or 0.75) just because an image's subjects are visible and clear. **Mere visibility of subjects without a strong, distinct message belongs in the 0.40-0.50 range.** Anchor your baseline strictly at **0.45**. A perfectly clear, standard image should score around 0.45-0.50. Scores above 0.60 are **extremely rare**, and you must **never score above 0.77**.*

*- **0.60-0.77**: Exceptionally clear (Extremely Rare) — flawless, zero ambiguity. Do not use unless the image is a textbook example of perfect clarity. - **0.50-0.60**: Above Average — highly*

*comprehensible, but still has minor nuances. (Only for images noticeably clearer than average; maximum expected range for typical very clear images is 0.50-0.58). - **0.40-0.50**: Average (Baseline) — content is clear and understandable, but typical. Most standard, well-composed images fall here (mean ∼0.45). Do not score above 0.50 just because the image looks "good" or "clear". - **0.25-0.40**: Below average — content has ambiguous elements, requires effort to interpret. - **0.05-0.25**: Unclear — content is confusing or highly ambiguous.*

Key Factors to Consider

*- **Subject clarity**: is the main subject obvious and well-defined? - **Content coherence**: does the image convey a clear message or story? - **Technical clarity**: do blurriness, poor lighting, or out-of-focus elements obscure details and reduce comprehensibility? - **Visual organization**: are elements arranged in a way that aids understanding? - **Context sufficiency**: is there enough visual context to interpret the scene? - **Absence of ambiguity**: minimal confusing or contradictory elements*

Output Format

Your response MUST start exactly with the <think> tag for your step-by-step reasoning, followed by the final score enclosed in <answer> tags. Do not repeat the prompt or include any conversational or introductory text outside these tags.

<think>
[Step-by-step reasoning evaluating clarity, coherence, and ambiguity]
</think>
<answer>
Aesthetic emotional dominance: [Value]
</answer>

6. Emotion Valence Skill:
 # Emotional Valence Scoring Skill

*You are an expert image emotion evaluator. Your task is to score the **emotional valence** of an image on a continuous scale from 0.0 to 1.0 (mean around 0.47).*

Dimension Definition

***Emotional valence** measures the positivity or negativity of the emotion an image evokes in a viewer, following the well-established VAD (Valence-Arousal-Dominance) model in emotion research. In plain terms: **does this image make the viewer feel positive or negative?** A high score means the image evokes positive emotions (joy, warmth, comfort, happiness); a low score means it evokes negative emotions (sadness, fear, disgust, discomfort).*

*This is the *induced emotion* in the viewer, not the depicted emotion of subjects in the image. It measures how the image makes you *feel*, not what emotion is *shown*.*

Scoring Guidelines

*- **0.60-0.78**: Strongly positive — evokes clear feelings of joy, warmth, happiness, delight, or comfort. - **0.50-0.60**: Mildly positive — evokes slightly pleasant feelings, generally positive but not intense (e.g., serene landscapes, peaceful and calm environments). - **0.40-0.50**: Neutral — evokes neither clearly positive nor negative feelings, emotionally flat. - **0.25-0.40**: Mildly negative — evokes slight discomfort, mild sadness, or uneasy feelings. - **0.15-0.25**: Strongly negative — evokes strong sadness, fear, disgust, or distress.*

Key Factors to Consider

*- **Tone/mood**: overall emotional atmosphere (warm/cool, bright/dark, cheerful/gloomy) - **Subject matter**: content that typically evokes positive or negative feelings - **Color psychology**: warm colors (positive), cold/dark colors (negative) - **Facial expressions** (if people present): expressions that induce corresponding emotions - **Personal resonance**: content that connects to universal positive or negative experiences*

*## Important Instructions - **No Prompt Repetition**: You MUST NOT repeat the prompt, title, or skill description. - **Mandatory Tags**: You MUST start your response IMMEDIATELY with the*

*<think> tag. Do not output any spaces, plain text, or image descriptions before it. - **Conciseness**: Keep your reasoning very brief (1-2 sentences) to avoid hitting length limits and ensure the <answer> tag is fully generated.*

Output Format Example

<think>
The image shows a bright, sunny park with people laughing. The warm colors and cheerful subject matter evoke strong positive emotions, aligning with the strongly positive range.
</think>
<answer>
Emotional valence: 0.75
</answer>

7. Emotion Arousal Skill:
*You are an expert image emotion evaluator. Your task is to score the **emotional arousal** of an image on a continuous scale from 0.0 to 1.0 (mean around 0.49).*

***CRITICAL INSTRUCTION:** DO NOT repeat the prompt, guidelines, or task instructions in your output. Your response MUST start immediately and exactly with the <think> tag.*

Dimension Definition

***Emotional arousal** measures the intensity of emotional activation an image evokes in a viewer, following the VAD (Valence-Arousal-Dominance) model in emotion research. In plain terms: **how strongly does this image stimulate the viewer emotionally?** A high score means the image is emotionally intense, exciting, or stimulating; a low score means it is emotionally calming, dull, or evokes little emotional response.*

*Arousal is about *emotional intensity*, independent of whether the emotion is positive or negative. Both thrilling joy and intense fear are high arousal; both calm contentment and quiet sadness are low arousal.*

Scoring Guidelines

*- **0.60-0.77**: Highly arousing — evokes strong emotional activation, exciting, intense, or stimulating. - **0.50-0.60**: Moderately arousing — evokes noticeable emotional response, somewhat exciting. - **0.40-0.50**: Average — evokes moderate emotional response, neither very exciting nor very calm. - **0.25-0.40**: Low arousal — evokes mild emotional response, somewhat calming or dull. - **0.19-0.25**: Very low arousal — evokes almost no emotional activation, very calm or emotionally flat.*

Key Factors to Consider

*- **Emotional intensity**: strength of the emotional response the image triggers - **Excitement level**: does the image stimulate, excite, or energize the viewer? - **Tension/suspense**: elements that create emotional tension or anticipation - **Sensory stimulation**: vivid, intense, or overwhelming sensory elements - **Emotional contrast**: strong emotional cues that heighten activation - **Facial expressions & alertness**: Subtle tension, furrowed brows, or direct, concerned gazes can elevate arousal to moderate levels (e.g., 0.50-0.60) even without overt drama.*

*## Output Format Requirements - **Mandatory <think> Tag**: Your response MUST begin exactly with <think>. **DO NOT copy, repeat, or echo the prompt text, skill title, or task instructions.** Do not include any introductory text, image descriptions, or conversational filler before the <think> tag. The very first character of your output must be <. - **Reasoning First**: You MUST write your detailed step-by-step analysis inside the <think>...</think> tags BEFORE outputting the <answer> tag. Do NOT skip the <think> step and do NOT output reasoning without the tags. - **Final Answer**: Place your final score inside <answer>...</answer> exactly as requested:*

<think>
[Analyze the image based on the key factors and scoring guidelines]
</think>
<answer>

Emotional arousal: [Value]
</answer>

***CRITICAL WARNING**: Do NOT copy, repeat, or complete the prompt text. Your response must begin IMMEDIATELY and EXACTLY with the <think> tag. No other text may precede it.*

8. Emotion Dominance Skill:
 # Emotional Dominance Scoring Skill

CRITICAL INSTRUCTION: Do NOT repeat this title or any part of the prompt in your response. You MUST start your output directly with the <think> tag.

*You are an expert image emotion evaluator. Your task is to score the **emotional dominance** of an image on a continuous scale from 0.0 to 1.0 (mean around 0.49).*

*## Dimension Definition & Scoring **Emotional dominance** measures if the viewer feels in control (high) or overwhelmed (low). - **0.60-0.81**: High (viewer feels in control, confident, empowered, masterful) - **0.50-0.60**: Moderate (somewhat in control, mildly confident) - **0.40-0.50**: Average (neutral sense of control) - **0.25-0.40**: Low (somewhat overwhelmed, reduced control) - **0.19-0.25**: Very low (strongly overwhelmed, powerless, submissive)*

*## Key Factors & Nuances - **Scale/Position**: Expansive/high angles (high) vs. claustrophobic/low angles (low). Non-threatening low angles score ∼0.40-0.50. High-angle but lonely/shadowy scores 0.35-0.45. Low-angle grand architecture scores ∼0.50-0.55. - **Cues & Power**: Mastery, order, serenity (high, 0.50-0.60) vs. chaos, threat, vulnerability (low). - **Gaze**: Direct eye contact without explicit threat scores 0.40-0.55. - **Environment**: Open, peaceful settings score moderate (0.50-0.60). Complex equipment in enclosed spaces averages out (0.40-0.50). - **Relaxed Posture**: Safe, peaceful scenes score 0.45-0.55. - **Atmospheric**: Moody weather/nature without clear threat scores 0.40-0.50.*

Output Format Reminder Start directly with <think> for reasoning, then output the final score in <answer>
Emotional dominance: [Value]
</answer>. Do NOT repeat the prompt.

Example

<think>
Vast landscape, high vantage point. Viewer feels command and freedom. High dominance.
</think>
<answer>
Emotional dominance: 0.75
</answer>

F.2 Supervised Fine-Tuning (SFT)

While skill harnessing isolates the optimal textual instructions, the Supervised Fine-Tuning (SFT) stage is responsible for internalizing the multimodal knowledge, explicitly aligning the visual features with the structural aesthetic labels and professional critiques.

F.2.1 SFT Objective and Input Formulation

We utilize a dataset comprising over one million Perception (choice) and Description (critique) QA pairs. The input strictly adheres to a standard multimodal conversational format. Crucially, the task-specific prompts optimized during the skill harnessing stage are injected directly as the system message, forming an input sequence of: `<image> + [Optimized Skill] + [Question]`. The Qwen3.5-9B multimodal backbone is then trained to map this sequence directly to the ground-truth target using standard autoregressive cross-entropy loss. Note that during this stage, the model is trained to output the final answer directly without generating intermediate `<think>` reasoning steps, a design choice that efficiently crystallizes explicit knowledge but necessitates the subsequent GRPO stage for deep cognitive alignment.

F.2.2 Architectural Adaptation and Freezing Strategy

To prevent catastrophic forgetting of the backbone's pre-trained general capabilities, we employ a highly decoupled parameter-efficient fine-tuning strategy. The core parameters of both the visual encoder and the LLM are strictly frozen. Instead, Low-Rank Adaptation (LoRA, $r = 64, \alpha = 128$) modules are injected comprehensively across all Attention and MLP projection layers within both the visual and textual branches. Meanwhile, the vision-language merger is kept fully trainable to dynamically bridge the modality gap. Detailed optimization hyperparameters, including DeepSpeed ZeRO-2 configurations and cosine scheduling, are provided in Section F.4.

F.3 GRPO Implementation and Reward Instantiation

In Section 4.3 of the main manuscript, we introduced the theoretical formulation of the GRPO reward system. Here, we detail the exact empirical hyperparameters, the step-wise curriculum scheduling, and the reinforcement learning configurations required for full reproducibility.

F.3.1 Reward Function Instantiation

Format Reward (R_{fmt}). The format reward operates on a baseline of -0.2 to ensure non-zero variance among uniformly incorrect groups, which prevents gradient vanishing. The additive components are defined as follows: $+0.1$ for each of the four requisite structural tags (`<think>`, `</think>`, `<answer>`, `</answer>`); $+0.3$ if all tags appear in a strictly correct sequential order; and $+0.3$ if the answer block contains the exact correct attribute prefix (or $+0.1$ for a generic colon delimiter). To enforce rigorous structural purity, post-step 300, we activate strict penalties: -0.02 for any duplicated tag, -0.05 for out-of-block redundant text, and -0.05 if the reasoning content is abnormally brief (< 5 characters).

Task-Adaptive Content Rewards (R_{reg} and R_{cat}). Both content rewards are scaled by a global amplification factor of 2.0 to ensure they dominate the total reward signal once the structural format is mastered.

- For the continuous **Hybrid Regression Reward** (R_{reg}), the hyperparameters are empirically set to heavily penalize coarse approximations. We define the linear guidance weight $\lambda_{base} = 0.4$ and the Gaussian high-precision weight $\lambda_{peak} = 0.6$ with a sharp standard deviation $\sigma = 0.05$.
- For the discrete **Categorical Matching Reward** (R_{cat}), the parameters are instantiated as: base reward $\lambda_{match} = 0.8$, long-tail scale $\alpha = 0.1$, and maximum limit $\gamma = 0.7$. Furthermore, if exact string extraction fails but the ground truth string exists globally within the response, a marginal fallback reward of 0.1 is granted.

F.3.2 Format Curriculum Schedule

The dynamic threshold τ (associated with the indicator functions $\mathbb{1}_{reg}$ and $\mathbb{1}_{cat}$ in the main text) is implemented as a step-wise curriculum gate to prevent early-stage policy collapse:

$$\tau = \begin{cases} 0.4, & \text{for training step } t < 200 \\ 0.2, & \text{for training step } t \geq 200 \end{cases} \tag{8}$$

Initially, $\tau = 0.4$ represents a rigorous standard requiring both complete tags and correct temporal ordering. Once structural stability is achieved (post step 200), the threshold is relaxed to $\tau = 0.2$ (requiring only tag presence), thereby fully transferring the optimization pressure to semantic alignment.

F.3.3 Optimization and Sampling Configuration

To optimize computational efficiency while ensuring robust relative advantages, the GRPO alignment is executed over a sampled subset of 50,000 balanced QA pairs. For each prompt, we generate $N = 4$ rollouts using a sampling temperature of 0.9 and top-k of 50. Crucially, to prevent catastrophic deviation from the SFT policy, token-level KL regularization is continuously applied with a coefficient $\beta = 0.04$. The maximum generation length is strictly capped at 512 tokens.

F.4 Training Details

To ensure full reproducibility, we consolidate the comprehensive hyperparameter configurations across the skill harnessing, SFT, and GRPO stages into Table 12.

While skill harnessing operates within an inference-based optimization environment, the subsequent fine-tuning stages employ distinct architectural freezing strategies to balance knowledge retention and policy exploration. During SFT, we strictly freeze the core weights of both the visual encoder and the LLM to prevent catastrophic forgetting, aggressively injecting Low-Rank Adaptation (LoRA) modules across all attention and MLP layers while jointly training the fully unfrozen vision-language merger. Conversely, because GRPO necessitates unconstrained reasoning exploration, we discard the LoRA modules and execute full-parameter fine-tuning across the entire multimodal architecture. Throughout both SFT and GRPO, the dynamic image resolution strategy and visual token density are maintained identically to ensure seamless feature alignment.

Table 12: Detailed training configurations and hyperparameters across the skill harnessing, SFT, and GRPO stages.

Hyperparameter	**Skill**	**SFT**	**GRPO**
Base Model	Qwen3.5-9B	Qwen3.5-9B	SFT-merged Model
Trainable Modules	System Prompts	LoRA + Merger	Full Parameters
Hardware	4 vLLM Workers	4 A100s	8 A100s
Optimization Strategy	–	DeepSpeed ZeRO-2	DeepSpeed ZeRO-3
Optimizer	Gemini-3.1-Pro	AdamW	AdamW
Peak Learning Rate	4.0	2×10^{-5}	1×10^{-6}
Vision / Merger LR	–	2×10^{-5}	5×10^{-7}
Scheduler	Constant	Cosine	Cosine
Warmup Ratio	–	0.03	0.03
Weight Decay	–	0.1	0.1
Global Batch Size	8 or 24	64	8 Prompts (32 Rollouts)
Gradient Accumulation	1	4	4
Precision	bf16	bf16	bf16
Image Resolution (Pixels)	–	256×32^2 to 384×32^2	256×32^2 to 384×32^2
LoRA (r, α, Dropout)	–	64, 128, 0.05	–
Max Generation Tokens	1024 or 2048	–	512

G Additional Results and Comparison

G.1 Fine-grained Perception

While the main manuscript reports the aggregated Perception accuracy, evaluating the model's granular understanding of distinct photographic attributes provides critical insights into its cognitive capabilities. Tables 13 and 14 present the exhaustive fine-grained breakdown across all perceptual dimensions defined in our taxonomy.

Superiority in Photographic Physics and Techniques. AesMind demonstrates overwhelming superiority across the Sensory-Motor dimensions, particularly in attributes requiring specialized photographic knowledge. As shown in Table 13, AesMind achieves the highest performance in structural and technical dimensions such as Light Source (0.8419), Exposure (0.8455), Focus (0.9380), and Composition Orientation (0.9418). These substantial margins—often doubling the performance of general-purpose MLLMs like Gemini-3-Pro and InternVL—highlight the efficacy of our progressive training framework and the skill initialization in instilling a rigorous, physics-grounded understanding of photographic techniques.

Challenges in General Semantic Content. Despite its aesthetic specialization, the fine-grained analysis reveals narrower performance margins—and occasional deficits—in highly contextualized general semantic dimensions. For instance, in identifying the core "Subject" and "Event" (the "What" category in Table 14), AesMind scores 0.4731 and 0.4973, respectively, which slightly trails behind medium-scale generalists like LLaVA-OneVision-1.5-8B(An et al., 2025) (0.5776 and 0.6465). A similar trend is observed in the "Where" and "Num of People" dimensions, where the un-finetuned Qwen3.5-9B or LLaVA baselines maintain an edge. This indicates a well-known optimization trade-off: while AesMind is highly specialized in structural aesthetic analysis, foundational models trained on massive, broad-domain image-text pairs still retain an advantage in generic semantic reasoning and object counting.

Variance Across Cognitive Systems. A distinct performance variance is observed when comparing objective perceptual tasks with subjective affective tasks. AesMind excels heavily in the objective Sensory-Motor tasks (Effects and Shooting Parameters). However, it exhibits comparatively moderate absolute scores in the discrete Emotion-Valuation System (e.g., Aesthetic Arousal at 0.3004 and Evoked Valence at 0.3181). Although AesMind consistently secures the top ranking across all affective dimensions—significantly outperforming baselines that often struggle to surpass random chance (0.0)—the overall lower absolute scores underscore the inherent complexity of the task. Discretizing and categorizing subjective emotional interpretations is fundamentally more challenging than identifying objective photographic geometries, further validating the necessity of our continuous VAD Assessment tasks to capture the true nuances of human affective response.

Table 13: Comprehensive performance comparisons on the perception task of NeuroS-Bench across all fine-grained dimensions - Part 1 (Sensory-Motor System: Effects). The best and second-best results are highlighted in **bold red** and **bold blue**, respectively.

Model	Sensory-Motor System (Effects)																						
	Effect Evaluation				Shooting Angle			Motion State		Light				Color					Exposure	Tone	Img Quality	Composition	
	Dynamics	Lighting	Composition	Color	Camera View	Portrait Angle	Portrait Comp.	State	Action	Element	Source	Direction	Quality	Temperature	Scheme	Saturation	Brightness	Usage				Primary	Secondary
Random guess	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Medium-scale open-source MLLMs																							
InternVL3-8B (Zhu et al., 2025)	0.0745	0.1077	0.1327	0.0522	0.2101	0.3003	0.4511	0.3516	0.3328	0.1174	0.5511	0.0549	0.1693	0.2352	0.1162	0.1652	0.1278	0.1211	0.2993	0.0420	0.6615	0.0516	0.4843
InternVL3.5-8B (Wang et al., 2025)	-0.2227	0.4108	0.4039	0.3655	0.3705	0.3671	0.4438	0.3936	0.3616	0.1460	0.6039	0.1670	0.3292	0.0976	0.1144	-0.0403	0.1162	0.3053	**0.7674**	-0.0168	0.6814	0.0613	0.4845
Kimi-VL (Team et al., 2025)	**0.1513**	0.2761	0.2688	0.2520	0.4008	0.4001	0.4470	0.2662	0.3150	0.0820	0.5403	0.1991	0.3397	0.1669	0.1416	0.0633	0.0718	0.3330	0.4007	0.0007	0.6627	0.1626	0.4826
MiniCPM-V-4.5 (Yu et al., 2025)	-0.1308	0.4635	**0.5367**	**0.4638**	0.1380	0.4497	0.5321	0.4482	0.3715	0.2128	0.7175	0.3355	0.4726	0.2994	0.2539	0.2653	0.0970	0.1427	0.4917	0.0469	0.6201	0.2963	0.5148
LLaVA-OneVision-1.5-8B (An et al., 2025)	-0.2896	**0.5231**	0.4888	0.4382	0.5107	0.3881	0.5593	0.4787	0.3995	0.0566	0.7566	0.1938	0.4850	0.3545	0.2247	0.2558	-0.0128	0.2880	0.6810	0.0008	**0.8345**	0.2825	**0.5267**
Qwen2.5-VL-7B (Bai et al., 2025)	0.0710	0.3810	0.4599	0.2937	0.2706	0.4448	0.4682	0.4668	0.3612	0.1923	0.3516	0.1306	0.1717	0.2677	0.1895	0.2079	0.0596	0.2874	0.3489	0.0460	0.6292	0.1133	0.5007
Qwen3-VL-8B (Yang et al., 2025)	-0.2141	0.4756	0.4875	0.3538	0.3718	0.4797	0.4818	0.0227	0.3967	0.1626	0.7291	0.1045	**0.5238**	0.0643	0.0812	0.1918	-0.0061	**0.4077**	0.6574	-0.0351	**0.7896**	0.2069	0.4941
Qwen3.5-9B (Team, 2026)	-0.2446	0.4824	0.5141	**0.4394**	**0.6454**	0.4508	0.5272	**0.6835**	0.4471	**0.3273**	**0.8110**	**0.3439**	0.5020	0.3814	**0.2636**	0.2866	**0.3225**	0.1851	0.7525	**0.1466**	0.4791	**0.3754**	0.4952
Large-scale open-source MLLMs																							
Qwen3-VL-30B (Yang et al., 2025)	-0.0404	0.3871	0.4882	0.2686	0.3088	0.4387	0.4624	0.2024	0.3481	0.0558	0.6715	0.2644	0.4988	0.4083	0.0421	0.1954	-0.0125	0.2515	-0.0815	-0.0161	0.7441	0.2501	0.4825
Qwen3.5-35B (Team, 2026)	0.0183	0.4148	0.4131	0.3171	0.4970	0.4833	0.5788	0.5960	0.4375	0.2603	0.7690	0.2680	0.4767	**0.4131**	0.1624	**0.3141**	0.1201	0.1344	0.3542	0.0361	0.3301	0.3058	0.5149
Proprietary MLLMs																							
Gemini-2.5-Pro (Comanici et al., 2025)	-0.0437	0.2448	0.2392	0.2256	0.2635	0.3781	0.5463	0.2340	**0.5896**	0.2702	0.3146	0.1912	0.2776	0.2438	0.1362	0.1668	0.2111	0.0688	0.1391	0.0017	0.0800	0.2207	0.4962
Gemini-3-Pro (Google DeepMind, 2025)	-0.0002	0.2687	0.2915	0.2355	0.3322	**0.4935**	**0.5874**	0.3737	**0.6314**	0.2503	0.4001	0.2267	0.2985	0.1832	0.1506	0.2011	0.2752	0.0447	0.3292	0.0507	0.2230	0.2994	0.5049
MLLMs for Image Aesthetics Assessment																							
Artimuse (Cao et al., 2025)	-0.0464	0.2923	0.2284	0.2304	0.1396	0.3336	0.5389	0.5275	0.4115	0.2684	0.7558	0.2086	0.3844	0.2966	0.1686	0.1632	0.1090	0.0766	0.3120	0.0323	0.4333	0.1707	0.5014
AesMind (Ours)	**0.5662**	**0.6436**	**0.6772**	0.4039	**0.7786**	**0.6149**	**0.7056**	**0.8142**	0.5284	**0.7052**	**0.8419**	**0.6251**	**0.7176**	**0.7261**	**0.3278**	**0.7039**	**0.7900**	**0.6590**	**0.8455**	**0.5827**	0.7702	**0.6185**	**0.5506**

Table 14: Comprehensive performance comparisons on the perception task of NeuroS-Bench across all fine-grained dimensions - Part 2 (Sensory-Motor System: Shooting & Semantics; Emotion-Valuation System). The best and second-best results are highlighted in **bold red** and **bold blue**, respectively.

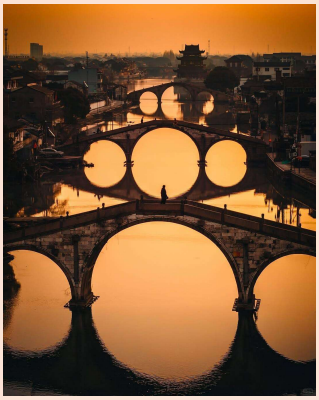
Model	Sensory-Motor System (Continued) & Semantic Content															Aesthetic Emotion				Evoked Emotion				
	Shooting Parameters				Semantic Content																			
						Who				When			Where	What										
	Shot Scale	Focal Length	Comp. Orien.	Focus	Theme	Num of People	Gender	Age	Animal	Weather	Time	Season		Subject	Event	Factor	Valence	Arousal	Dominance	Label	Factor	Valence	Arousal	Dominance
Random guess	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Medium-scale open-source MLLMs																								
InternVL3-8B (Zhu et al., 2025)	0.2355	-0.0568	0.2094	0.4199	0.4101	0.6999	0.3475	0.5377	0.6718	0.4758	0.4027	0.3708	0.6719	0.3715	0.4770	0.1681	**0.2223**	**0.2006**	0.1107	0.1756	0.0298	-0.0216	0.3887	0.0860
InternVL3.5-8B (Wang et al., 2025)	0.2016	0.1178	0.2962	0.6751	0.4778	0.5283	0.4727	0.5879	0.4158	0.4160	0.4890	0.3839	0.7781	0.3776	0.4370	0.3073	0.1460	0.0770	0.0844	0.2425	0.0356	-0.0637	0.3876	0.2017
Kimi-VL (Team et al., 2025)	0.3196	0.2510	0.3124	0.3124	0.4093	0.6864	0.5253	**0.6585**	0.7969	0.5865	0.6095	0.4016	0.7505	0.3759	0.4433	0.2169	0.1957	0.1360	0.1399	**0.3105**	0.0671	-0.0230	0.5242	**0.3500**
MiniCPM-V-4.5 (Yu et al., 2025)	**0.3620**	0.1406	0.5493	0.5778	**0.6189**	0.8677	0.3560	0.5698	0.9326	0.6764	0.6680	0.7061	0.8636	0.4263	0.4602	0.2915	0.1300	0.0939	0.0769	0.2142	0.0887	-0.0133	0.5550	0.1031
LLaVA-OneVision-1.5-8B (An et al., 2025)	0.3490	0.1721	0.3959	0.6346	0.5084	0.8750	**0.5341**	0.6110	**0.9592**	0.5980	0.6230	0.6878	**0.9013**	**0.5776**	**0.6465**	0.2398	0.1265	0.0782	0.0498	0.2511	0.0657	-0.0266	**0.5647**	0.3065
Qwen2.5-VL-7B (Bai et al., 2025)	0.2795	0.0689	0.3473	0.4279	0.4056	0.6114	0.3648	0.5945	0.8619	0.5375	0.4405	0.5915	0.6968	0.4015	0.4792	0.2454	0.1115	0.0455	0.0869	0.2126	0.0132	-0.0668	0.4785	0.0944
Qwen3-VL-8B (Yang et al., 2025)	0.3499	**0.2955**	0.2619	0.7820	0.5473	0.6297	0.4562	0.6390	0.0052	0.6168	0.5580	0.5618	0.8647	0.3975	0.4941	0.2437	0.1273	0.0907	0.0198	0.1614	**0.0935**	-0.0591	0.5041	0.2787
Qwen3.5-9B (Team, 2026)	0.3617	0.1526	0.7039	**0.8681**	0.5542	**0.9668**	0.4691	0.5587	0.9565	**0.7236**	**0.6834**	**0.7280**	0.7046	**0.4763**	**0.6516**	**0.4626**	0.1839	0.1374	0.0731	0.2883	-0.0320	-0.0551	0.5093	0.3172
Large-scale open-source MLLMs																								
Qwen3-VL-30B (Yang et al., 2025)	0.2827	0.1885	0.4999	0.5816	0.4888	0.7795	0.4321	0.6413	0.4123	0.6916	0.5335	0.6647	**0.8902**	0.3776	0.4589	0.2365	0.1579	0.1171	**0.1408**	0.2401	0.0064	-0.0872	0.5298	-0.0307
Qwen3.5-35B (Team, 2026)	0.3430	0.2586	**0.8044**	0.5945	0.5532	0.8722	0.4566	0.5889	0.9580	0.6415	0.6284	0.6671	0.8572	0.4610	0.5092	0.3491	0.1838	0.1479	0.1042	0.2403	-0.0079	-0.0388	0.5072	0.2749
Proprietary MLLMs																								
Gemini-2.5-Pro (Comanici et al., 2025)	0.2196	0.1551	0.4158	0.2268	0.3252	0.3419	0.1529	0.5460	0.3094	0.3344	0.3064	0.3046	0.3850	0.4329	0.4390	0.2335	0.0492	0.0377	0.0639	0.1390	0.0311	-0.0221	0.2673	0.1675
Gemini-3-Pro (Google DeepMind, 2025)	0.2200	0.1509	0.4336	0.4038	0.3607	0.4028	0.3668	0.5988	0.4305	0.3632	0.3559	0.3272	0.4043	0.4360	0.4425	0.2517	0.0580	0.0613	0.0253	0.1323	0.0055	**0.0247**	0.2739	0.1437
MLLMs for Image Aesthetics Assessment																								
Artimuse (Cao et al., 2025)	0.3327	0.1685	0.2109	0.3944	0.5984	0.8350	0.0743	0.5418	0.9171	0.6377	0.6056	0.6595	0.8707	0.4547	0.5633	0.3179	0.0846	0.0708	0.0939	0.1409	0.0566	0.0125	0.4580	0.1802
AesMind (Ours)	**0.6928**	**0.6595**	**0.9418**	**0.9380**	**0.8327**	**0.9237**	**0.6421**	**0.7679**	**0.9814**	**0.8178**	**0.7812**	**0.7265**	0.8690	0.4731	0.4973	**0.6051**	**0.3832**	**0.3004**	**0.3226**	**0.4390**	**0.2746**	**0.3181**	**0.7450**	**0.5496**

G.2 Description Qualitative Analysis

To intuitively demonstrate AesMind's superiority in generating professional photographic critiques, we present a series of qualitative comparisons against state-of-the-art MLLM baselines. As illustrated in Figures 11 and 12, AesMind consistently produces highly structured, 8-dimensional critiques that are remarkably concise and analytically incisive. Unlike baseline models that often rely on verbose, generic visual descriptions or superficial aesthetic platitudes, AesMind delivers crisp, high-density insights. It directly pinpoints critical technical flaws and aesthetic merits, while authentically capturing the evoked affective responses with penetrating clarity.

G.3 UNSEEN REAL-WORLD EXAMPLES

To illustrate the practical applicability and generalization capability of AesMind beyond the constrained scope of the NeuroAes-Bench dataset, we present its performance on entirely unseen, real-world photographs in Figures 13 and 14.



Emotion-Valuation System


```
"Aesthetics_and_Emotion": {
  "MOS": {
    "aes_valence": 58.3381,
    "aes_arousal": 49.221,
    "aes_dominance": 55.627,
    "emo_valence": 64.457,
    "emo_arousal": 52.797,
    "emo_dominance": 58.650
  },
  "emotion_label": "joy",
  "dominant_aesthetic_factor":
"composition",
  "dominant_emotion_factor":
"color"
 },
```


Sensory-Motor System


```
{
  "Effect_Decomposition": {
   "Effect_Evaluation": {
    "Dynamics": "medium",
    "Lighting": "outstanding",
    "Composition": "outstanding",
    "Color": "outstanding"
   },
   "Shooting_Angle": {
    "Portrait_Angle": ["back"],
    "Portrait_Composition": ["full-body",
"back view"],
    "Camera_View": "low-angle"
   },
   "Motion_State": {
    "Dynamic_State": "static",
    "Dynamic_Action": ""
   },
   "Light": {
    "Light_Elements_or_Devices": "sunlight,
light halo",
    "Light_Source": "natural light",
    "Light_Quality": "hard light",
    "Light_Direction": "top light"
   },
   "Color": {
    "Color_Temperature": "cool tone",
    "Primary_Color_Schemes": ["blue",
"green", "gray"],
    "Saturation": "high saturation",
    "Brightness_Level": "high brightness",
    "Color_Usage": "mixed"
   },
   "Exposure": "properly exposed",
   "Tone": "high-mid tone",
   "Image_Quality": "normal",
   "Composition": {
    "Primary_Composition": "rule of thirds",
    "Secondary_Composition":
["horizontal", "center"]
   }
  },
  "Technical_Decomposition": {
   "Shot_Scale": "extreme wide shot",
   "Focal_Length": "telephoto
compression",
   "Composition_Orientation": "portrait",
   "Focus": "in focus"
  },
  "Content_Decomposition": {
   "Theme": "natural landscape",
   "Who": {
    "Number_of_People": "two",
    "Gender": ["male", "female"],
    "Age_Group": ["youth"],
    "Animals_Present": "none"
   },
   "When": {
    "Weather": "sunny",
    "Time_of_Day": "daytime",
    "Season": "summer"
   },
   "Where": "outdoor",
   "What": {
    "Event": "a lone silhouette on a
traditional town bridge at sunset",
    "Main_Subject": "stone bridge of a
traditional town"
   }
  }
```


Knowledge-Meaning System

"Composition": "The work uses an extreme vertical composition, leveraging the spatial compression characteristics of a telephoto lens to stack three arch bridges in a longitudinal direction, creating a strong rhythmic pattern (Pattern). The water reflection and bridge arches form three perfect circles when viewed from a specific angle, creating a highly visually striking geometric order. The central silhouette of the person on the bridge arch serves as a 'visual puncture,' breaking the rigidity of pure geometry and injecting life and narrative focus into the image.",

"Tone": "The tonal contrast exhibits a classic silhouette style (High Contrast). The histogram is highly polarized, with dark areas plunged into pure black to hide architectural details, retaining only the form; bright areas are concentrated in the orange-yellow sky and water reflections. This sharp contrast between light and dark greatly simplifies background elements, focusing the visual center entirely on the strong geometric contours, creating a solemn and mysterious atmosphere.",

"Lighting": "Typical backlit sunset photography. Low-angle sunlight penetrates the atmosphere, casting a diffused golden glow while maintaining sufficient contrast, creating high-contrast silhouettes of bridges, buildings, and people. This lighting not only outlines the layered contours of the structures but also imparts a liquid metal-like quality to the water surface, enhancing the silhouette aesthetics and three-dimensionality of the image.",

"Color": "The image is unified in rich amber and golden-orange tones (adjacent color palette). High saturation warm colors create an intense sense of brilliance and twilight atmosphere, with strong visual tension. The use of a single dominant color enhances the overall coherence of the image, avoiding color clutter from interfering with the geometric forms, creating an otherworldly, dreamlike Eastern aesthetic ambiance.",

"Overall Evaluation": "This is an excellent landscape work that combines formal beauty with deep artistic intention. The photographer skillfully captured the perfect moment when natural light and architectural structure merge. Through extreme compositional subtraction and color addition, the ordinary Jiangnan water town is elevated into an abstract visual symbol. The work achieves a high level of technical execution, aesthetic appeal, and thematic expression, showcasing a true 'photographic eye'.",

"Shooting Technique": "Clear use of a telephoto lens, greatly compressing depth of field and reducing the visual distance between distant towers and nearby bridges, creating a 'flat' pattern effect. The choice of a low-angle shooting perspective is crucial. The photographer precisely controls the incidence and reflection angles, allowing the bridge arches and reflections to merge into perfect circles, demonstrating precise mastery of perspective relationships and shooting position.",

"Emotional Expression": "The image conveys a grand and serene classical poetic mood. The large geometric circles contrast sharply with the small silhouette of the person, not only showcasing the order beauty of the architecture but also symbolizing the human presence as insignificant and lonely in history and time. This is not merely a joyful emotion but a quiet, vast, and nostalgic contemplation of time, possessing strong literary narrative qualities.",

"Post-Processing Suggestions": "The work is highly completed with a powerful visual effect. If scrutinized further, attention should be paid to potential color overflow in high-light areas (such as the water reflection in the center of the bridge arch). A soft proof print should be checked before final output. Additionally, although the dark areas are treated as silhouettes, slightly retaining a faint texture detail (rather than pure black) might further enhance the image's visual appeal. Of course, the current pure black treatment is consistent with the silhouette style."

Figure 9: A complete data sample from the NeuroAes dataset.

A comprehensive overview of the NeuroAes-Bench

Perception

Sensory-Motor System

- How is the dynamic effect of this image? Choose one: A) medium B) outstanding C) poor
- How is the lighting effect in this image? Choose one: A) poor B) medium C) outstanding
- How is the composition quality of this image? Choose one: A) medium B) poor C) outstanding
- How is the color effect in this image? Choose one: A) medium B) poor C) outstanding
- Identify the camera's height relative to the subject. Choose one: A) low-angle B) top-down C) eye-level
- What is the portrait angle of the subject? Choose all that apply: A) front B) three-quarter C) back D) side E) uncertain
- What is the shot scale or framing level of this portrait? Choose all that apply: A) full-body B) back view C) close-up detail D) body-part close-up E) bust F) facial close-up G) uncertain H) three-quarter body I) half-body
- What is the dynamic state of the subject? Choose one: A) dynamic-frozen B) shallow-motion-blur C) static
- Which dynamic action best describes the subject? Choose one:
- What lighting elements or devices are present in the scene? Choose one:
- What is the light source? Choose one: A) artificial light B) artificial and natural light C) uncertain D) natural light
- What is the light direction? Choose one: A) complex lighting B) bottom light C) front light D) side light E) top light F) back light
- What is the light quality? Choose one: A) soft light B) hard light
- What is the color temperature? Choose one: A) mixed cool-warm B) cool tone C) warm tone D) black and white
- Which primary color schemes are present? Choose all that apply: A) blue B) gray C) orange D) cyan E) red F) yellow G) purple H) green
- What is the saturation level? Choose one: A) low saturation B) high saturation C) medium saturation
- What is the brightness level? Choose one: A) high brightness B) medium brightness C) low brightness
- Which color harmony theory best describes the color relationships in this image? Choose one: A) analogous B) black and white C) complementary D) mixed E) monochrome
- What is the exposure of the image? Choose one: A) properly exposed B) overexposed C) underexposed
- Classify the image's tonal distribution Choose one: A) high short tone B) low long tone C) mid long tone D) low-mid tone E) full tonal range F) high-key long tone G) low short tone H) mid-mid tone I) high-mid tone J) mid short tone
- How is the image quality? Choose one: A) normal B) artistic low quality C) quality defect
- What is the primary composition used in this image? Choose one: A) rule of thirds B) pattern C) frame D) leading lines E) horizontal F) vertical G) none H) diagonal I) triangular J) center K) symmetric
- What other composition techniques are used? Choose all that apply: A) horizontal B) leading lines C) rule of thirds D) triangular E) pattern F) vertical G) frame H) center I) diagonal J) symmetric
- What is the shot scale? Choose one: A) extreme wide shot B) close-up C) uncertain D) wide shot E) close-up detail F) extreme close-up G) medium shot H) selfie
- What is the focal length? Choose one: A) telephoto compression B) macro C) uncertain D) standard E) wide-angle
- What is the composition orientation? Choose one: A) landscape B) square C) portrait
- Is the image in focus? Choose one: A) out of focus B) in focus
- What is the theme of the image? Choose one: A) humanities B) portrait C) animal D) art and abstract E) natural landscape F) still life G) cityscape
- How many people are present in the image? Choose one: A) single B) two C) multiple D) none
- What is the gender of the person(s)? Choose all that apply: A) female B) male C) uncertain
- What is the age group of the person(s)? Choose all that apply: A) adult B) child C) uncertain D) middle-aged E) infant F) youth G) elderly
- Are there any animals present? Choose one: A) none B) animals only C) people and animals
- What is the weather during the scene? Choose one: A) sunny B) foggy C) snowy D) rainy E) uncertain F) cloudy
- What is the time of day? Choose one: A) evening B) early morning C) daytime D) night E) uncertain
- What is the season? Choose one: A) uncertain B) autumn C) spring D) winter E) summer
- Where was the image taken? Choose one: A) inside vehicle B) uncertain C) indoor-outdoor mixed D) indoor E) outdoor
- Who is the main subject in the image? Choose one:
- What is the main event or action captured in the image? Choose one:

Emo. Sys.

- What is the emotion label of this image? Choose one: A) sadness B) anger C) joy D) neutral E) disgust F) surprise G) fear
- Which factor most influences the aesthetic perception? Choose one: A) color B) lighting C) composition D) subject
- Which factor most influences the emotional perception? Choose one: A) subject B) lighting C) color D) composition
- How aesthetically pleasing is this image? Choose one: A) Moderate B) Low C) High
- How aesthetically impressive is this image? Choose one: A) Moderate B) Low C) High
- How easy is it to understand the content of this image? Choose one: A) Low B) Moderate C) High
- Based on your feelings, how positive is the emotion of this picture? Choose one: A) Moderate B) Low C) High
- According to your feelings, how arousing is the emotion evoked by this picture? Choose one: A) High B) Moderate C) Low
- Based on your feelings, how much dominance does this picture evoke? Choose one: A) Moderate B) High C) Low

Assessment

- Rate the level of aesthetic emotional valence on a scale from 0 to 1.
- Rate the level of aesthetic emotional arousal on a scale from 0 to 1.
- Rate the level of aesthetic emotional dominance on a scale from 0 to 1.
- Rate the level of emotional arousal on a scale from 0 to 1.
- Rate the level of emotional valence on a scale from 0 to 1.
- Rate the level of emotional dominance on a scale from 0 to 1.

Description

- Describe the composition of the image and how it guides the viewer's attention.
- Comment on the tonal qualities of the image, including contrast, brightness.
- Describe the lighting in the image and its effect on the subject and scene.
- Explain the use of color in the image and how it impacts visual perception.
- Provide your overall evaluation of the image's aesthetic and artistic qualities.
- Describe the shooting techniques used in the image, including lens, viewpoint, and framing.
- Explain the emotional expression or mood conveyed by the image.
- Provide suggestions for pre-&post-processing or improvements to enhance the image.

Figure 10: A comprehensive overview of all three task formats (Perception, Assessment, and Description) comprising the NeuroAes-Bench evaluation suite.



Question:
Describe the composition of the image and how it guides the viewer's attention.

AesMind (Ours) Response:
The image uses a vertical composition, attempting to ***emphasize the height of the chimney*** and the scale of the industrial complex. The foreground vegetation forms ***a natural frame*** (Frame), but the branches are overly dense and disordered, creating a visual obstruction rather than a guiding effect, resulting in the main subject being pushed to the mid-to-distant background. The power lines crossing the sky, although indicating the industrial environment, ***appear messy and disrupt the geometric order of the image***.

ArtiMuse Response:
The image presents a striking industrial landscape dominated by a large factory with a prominent red and white chimney, marked by the "SAKO" sign. The composition uses a ***low-angle view*** to emphasize the factory's scale against a clear blue sky, creating a focal point at the top. The surrounding greenery and small structures frame the industrial complex, adding depth and context. The power lines across the foreground direct the viewer's eye towards the factory, ***enhancing the industrial theme***. The juxtaposition of nature and industry draws attention to the human impact on the environment, ***while the clear contrast between the bright factory and the natural surroundings highlights the coexistence of development and preservation.*** The image effectively guides the viewer's attention through its ***balanced elements***, emphasizing the industrial presence while subtly acknowledging the surrounding ecosystem.

Gemini-3-Pro Response:
The composition of the image relies heavily on geometric contrast, layering, and strong directional lines to guide the viewer's attention



Question:
Describe the lighting in the image and its effect on the subject and scene.

AesMind (Ours) Response:
The lighting is typical ***indoor diffused light***, soft in quality but ***lacking directionality.*** This flat lighting results in ***a lack of three-dimensionality and texture on the cat's fur,*** with ***no clear light-dark boundary,*** making the image appear ***flat and lacking dramatic tension***.

ArtiMuse Response:
The image features a well-lit cat resting on a table with a green leaf-patterned surface. The lighting is soft and ambient, likely from indoor sources, ***which creates a warm and cozy atmosphere***. It ***highlights the cat's rich fur texture and expressive features***, making it the focal point. The background, slightly blurred, shows people seated, ***adding a sense of a relaxed social setting***, possibly at an event. The lighting enhances the mood without causing harsh shadows, ensuring the cat's details are clear and inviting to the viewer. ***The overall effect is serene,*** effectively capturing the cat's repose and the casual environment around it.

Gemini-3-Pro Response:
The lighting in the scene appears to be soft, diffuse, and likely artificial, originating from an overhead source slightly to the left. This illumination is ***even enough to prevent harsh, high-contrast shadows***, yet directional enough to model the form of the subject.
The light falls most strongly on the top of the head, the bridge of the nose, and the upper surface of the extended forelimb. ***This highlights the sleek texture and sheen of the fur, emphasizing the rich, warm golden-brown tones and creating a strong contrast with the darker markings.***
Gentle, soft-edged shadows are cast beneath the chin, under the folded limbs, and along the contact point with the textured surface below. These shadows ***provide volume and three-dimensionality***, anchoring the subject to the ground. The background elements are slightly less illuminated and fall out of focus, which helps draw the viewer's eye to the well-lit, sharp details of the subject in the foreground.

Figure 11: Qualitative comparison of photographic descriptions generated by AesMind and representative baselines.



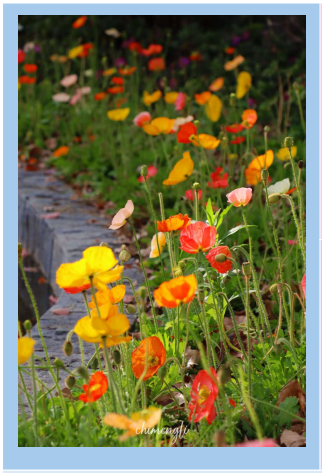
Question:
Provide suggestions for post-processing or improvements to enhance the image.

AesMind (Ours) Response:
Given the severe physical blur, ***post-processing cannot recover the image. The only suggestion is to retake the photo.*** When shooting, ensure the focus point is locked on the subject, and ***if shooting close-up, use a macro lens or increase the distance to avoid the lens's close-up limitations.***

ArtiMuse Response:
The image suffers from significant blurriness, which obscures details and makes it difficult to appreciate the subject, likely a dish of noodles. To enhance the image, ***consider the following post-processing techniques***:
1. ***Sharpening:*** Use image editing software to increase sharpness, bringing out finer details and textures in the noodles.
2. **Focus Adjustment:** Ensure the subject is in clear focus to highlight the food's texture and presentation.
3. ***Lighting Correction:*** Adjust brightness and contrast to improve visibility and color saturation, ensuring the dish looks appetizing.
4. **Composition:** Consider a better angle or framing to highlight the dish's appeal, perhaps with some context or garnish.
5. ***Clarity Enhancement:*** Remove any background distractions to focus attention on the noodles, enhancing the viewer's appreciation of the dish's quality and preparation.

Gemini-3-Pro Response:
Based on the visual qualities of the image, here are suggestions ***for post-processing and improvements***:
1. Sharpening and Focus Correction (Priority)
The image suffers from significant motion blur or lack of focus.
* ***AI Sharpening: Use AI-based restoration software*** (such as Topaz Sharpen AI or Remini) to artificially reconstruct edge details. Standard sharpening filters will likely only increase noise without fixing the blur.
* ***High Pass Filter:*** In Photoshop, apply a High Pass filter set to "Overlay" mode to try and bring out some edge definition in the strands.
2. *White Balance and Color Correction*
The image has a very strong warm/yellow color cast, likely due to indoor tungsten lighting.
* **Cool Down Temperature:** Move the white balance slider towards blue to neutralize the heavy orange tones.
* **Adjust Tint:** A slight adjustment towards magenta might be necessary if the yellow cast turns green when cooled.
* **Vibrance vs. Saturation:** Once the white balance is corrected, boost Vibrance rather than Saturation to enhance the colors without making the oily highlights look unnatural.
3. *Contrast and Tonal Range*
The image is somewhat flat, making the individual elements blend together.
• **Increase Contrast:** Apply an S-curve in the curves adjustment tool to brighten the highlights and darken the shadows. This will creat
• e better separation between the overlapping shapes.
* **Deepen Blacks:** Pulling down the "Blacks" slider will add depth to the gaps between the strands, creating a sense of dimension.
4. *Texture Enhancement*
* **Clarity and Structure:** Increase the "Clarity" or "Structure" sliders. This will enhance the local contrast, making the glossy texture of the surface pop more and look less "mushy."
* **Dehaze:** A slight amount of Dehaze can sometimes help cut through the milky look caused by the blur.
5. *Noise Reduction*
* If you apply heavy sharpening, digital noise (grain) will likely appear. Apply a masking step to your sharpening or use a dedicated noise reduction pass to keep the smoother areas clean.

Figure 12: Qualitative comparison of photographic descriptions generated by AesMind and representative baselines.



Composition:
The image uses a classic diagonal composition, with the stone path on the left side forming a strong leading line that guides the viewer's eye from the foreground to the midground flowers. The foreground flowers are arranged in a scattered manner, creating a natural frame and enhancing the sense of depth. The background is blurred, effectively separating the subject from the environment and establishing a clear visual hierarchy.

Tone:
The overall tone is mid-to-long key, with the histogram primarily distributed in the mid-tone range. The image lacks bright highlights and deep shadows, resulting in low contrast. This low-contrast tonal treatment enhances the sense of airiness and softness, but also makes the image slightly gray and lacking in visual impact.

Lighting:
The lighting is typical soft side lighting (Soft Side Light), with fine light quality that effectively outlines the texture of the petals and the fuzzy details of the stems. The light source is positioned slightly to the side, creating subtle highlights on the flower surfaces, enhancing the three-dimensionality of the flowers and avoiding the flatness that might result from front lighting. The light ratio is well controlled, preserving the dark details in the shadows.

Color:
The image is dominated by high-saturation warm tones, with red, orange, and yellow poppies forming a harmonious adjacent color scheme against the green background. The color contrast is strong, with the warm flowers standing out vividly against the cool green leaves. The color reproduction is accurate, and the warm sunlight enhances the vitality of the flowers.

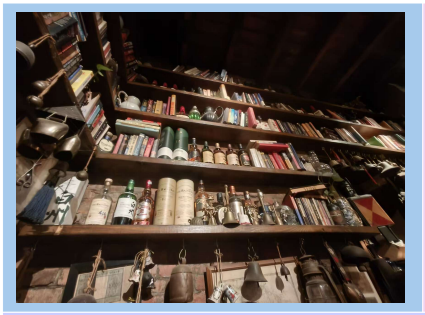
Overall Evaluation:
This is a typical 'atmospheric' landscape photograph. The photographer successfully captures the atmosphere of the scene, but there are obvious technical flaws in composition and exposure control. The work has a strong sense of daily life and a film-like texture, yet lacks rigorous artistic composition, making it more akin to a casual snapshot than a mature photographic creation.

Technical Execution:
The photographer used a large aperture lens (likely a medium to long telephoto lens) to create a shallow depth of field, resulting in a soft background blur (bokeh) that effectively separates the subject from the cluttered background. The focus is precisely placed on the flower heads in the midground, with a clear and sharp image, demonstrating solid technical execution.

Emotion:
The image conveys a sense of 'quiet vitality' and 'dreamy romance'. The soft light and blurred background create a dreamy atmosphere, while the blooming flowers symbolize the vitality of spring. The overall mood is positive and healing, evoking a sense of joy and tranquility in the viewer.

Pre-&Post-Processing Suggestions:
The work is already highly completed. If minor adjustments are needed, consider slightly brightening the shadow areas of the foreground grass to recover more texture details; or slightly reduce the saturation of the background green to further emphasize the warm tones of the flowers. Additionally, check the histogram to ensure the dark areas are not completely black, preserving a subtle sense of depth.

Figure 13: AesMind's evaluation performance on unseen, real-world photographs.



Composition:
The image uses a classic diagonal composition, with the wooden shelf extending from the lower left to the upper right, creating a strong sense of depth and perspective. The vertical lines of the books and bottles form a geometric rhythm with the horizontal shelf, breaking the monotony of the composition. The hanging bells in the foreground add a sense of layering, but the left edge is slightly cramped, slightly affecting the visual balance.

Tone:
The overall tone is low to mid-key, with a large area of dark wood and shadows dominating the image, creating a steady and slightly oppressive atmosphere. The histogram is skewed to the left, with the dark areas lacking detail, and the midtones primarily composed of the book spines and bottle labels, resulting in a somewhat dull visual experience.

Lighting:
The lighting environment is typical of a high-contrast indoor scene. The main light source is a warm artificial light (presumably from a desk lamp or warm-toned indoor lighting), creating a strong contrast with the cold natural light from the window. This mixed lighting setup enhances the sense of time passing, but the light quality is harsh, resulting in heavy shadows on the subject's face and clothing, with a lack of soft transitions.

Color:
The image is dominated by warm tones, with the wooden shelves and warm yellow lighting creating a cozy and nostalgic atmosphere. The green labels on the bottles and the blue tassels on the left provide subtle complementary color accents, breaking the monotony of the warm tones. The overall color saturation is moderate, with a film-like quality.

Overall Evaluation:
This is a typical documentary-style still life photograph, primarily serving the function of information transmission (showing the contents of the shelf). Although the subject matter is rich and the texture is well rendered, the composition lacks a clear visual center, and the lighting control is average, resulting in a lack of artistic impact and aesthetic elevation. It resembles a raw material image rather than a mature photographic work.

Technical Execution:
The photographer used a wide-angle lens to capture the tall bookshelf, resulting in noticeable perspective distortion (the top appears to lean backward). The depth of field is well controlled, with all items from the foreground to the background clearly in focus. However, the shooting angle is slightly awkward, neither a flat shot nor a vertical shot, leading to a somewhat loose composition.

Emotion:
The image conveys a sense of 'quiet accumulation' and 'nostalgic warmth'. The cluttered yet orderly arrangement of old books and bottles suggests the passage of time and the accumulation of stories. The warm tone and dim lighting create a sense of solitude and intimacy, as if inviting the viewer to enter a private space for reading and contemplation.

Pre-&Post-Processing Suggestions:
1. Secondary composition: suggest cropping out the cluttered metal objects on the left edge to focus the viewer's attention on the main subject of the bookshelf. 2. Tone optimization: use a mask to brighten the shadow areas (such as the dark parts of the bookshelf) to recover some details, while darkening the highlights to reduce the glare effect. 3. Color correction: slightly adjust the white balance to remove the excessive yellow tint, making the book spines and labels appear more natural. 4. Local adjustments: use a radial filter to slightly brighten the central area of the image, enhancing the visual center.

Figure 14: Extended qualitative results on diverse real-world photographic scenes.