AESCASE: TOWARDS INTERPRETABLE IMAGE AESTHETIC ASSESSMENT VIA BADCASE DISCOVERY

Anonymous authors
Paper under double-blind review

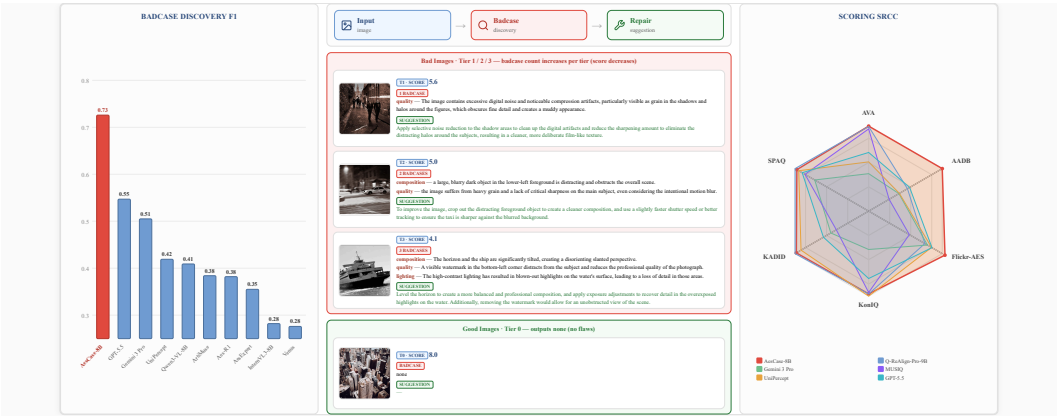

Figure 1: AesCase advances conventional aesthetic scoring to badcase discovery: given an image, AesCase-8B jointly outputs a score and grounded flaws (category + description). The middle shows tier examples (Tier 0 good images output *none*, Tiers 1–4 report up to 1–4 flaws). Left/right compare badcase-discovery F1 on AesCase-bench and scoring SRCC on six datasets.

ABSTRACT

Most existing image aesthetic assessment methods predict an overall score or preference, yet fail to answer where the problem lies, why it occurs, and how to improve it. To repair a flawed image, a model must first generate concrete repair suggestions, which in turn depends on precisely discovering and localizing its aesthetic flaws. To address this issue, we propose AesCase, a framework for aesthetic badcase discovery: 1) we propose a score-guided five-tier construction method that simulates the human badcase distribution to balance annotation efficiency and quality; building on it, we construct the AesCase-20k dataset and the AesCase-bench benchmark; 2) we train a baseline model AesCase-8B on mixed flaw-discovery and scoring data, jointly learning flaw discovery and score prediction, and generalize flaw localization into repair suggestions (suggestion). Experiments show that AesCase-8B achieves the best flaw-discovery F1 on AesCase-bench (0.726), while producing almost no false positives on good images (0.02 vs. GPT-5.5's 4.05), and maintains competitive scoring performance on six IAA/IQA benchmarks. Furthermore, this repair-oriented, interpretable feedback closes the loop from "where is the problem" to "how to fix it", making AesCase a plug-and-play model for image editing and aesthetic enhancement.

1 INTRODUCTION

Image aesthetic assessment has long centered on overall scores, preferences, or ratings. These paradigms answer "is this image good", but struggle to answer "where is it bad, why, and how should it be fixed". Users want actionable, interpretable feedback pointing to specific problems, such as whether the composition is imbalanced, the subject is occluded, the lighting is too harsh,

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.

the color is washed out, the background distracts from the subject, or a local fusion is unnatural. To repair a bad image, a model must first produce concrete repair suggestions; and to produce actionable suggestions, it must first precisely discover and localize the concrete aesthetic flaws (badcases): only after answering "where and why it is bad" can it answer "how to fix it". Badcase discovery therefore forms the foundation of repair-oriented feedback. Figure 1 illustrates the core paradigm: given an image, AesCase-8B jointly outputs an aesthetic score and concrete visible flaws (badcases), along with repair suggestions.

We summarize this capability as *badcase discovery*: the model should not only judge image quality, but also discover concrete aesthetic cases, give grounded problem descriptions, and, as far as possible, provide repair suggestions. This task sits between conventional aesthetic scoring and aesthetic critique, finer-grained than the former, and more structured than the latter. Compared with score-only models, badcase discovery is closer to real photography guidance, generative feedback, and image-editing workflows; compared with models that output only broad descriptions, it demands stronger visual-description and causal-explanation ability. Here "grounded" means the flaw description points to specific regions, objects, and causes (in natural language, rather than bounding boxes or pixel masks), making the feedback verifiable and able to directly drive editing models that take text instructions as input.

However, most existing work does not model this capability specifically. Traditional IQA/IAA methods focus on overall scores or quality prediction; recently emerging aesthetic-critique works (e.g., Cao et al. (2026), Du et al. (2026)) begin to generate fine-grained textual analyses, but mostly remain at the level of general commentary, lacking structured diagnosis centered on "badcase + suggestion" and thus being hard to directly drive image editing. More seriously, existing models commonly suffer from "nitpicking": they pile up many flaws even on high-quality good images (e.g., GPT-5.5 averages 4.05 false-positive flaws on good images), whereas a truly usable critique model should "not falsely flag" good images.

Motivated by this, we propose AesCase-bench and the score-guided data recipe used to construct it. AesCase-bench is a benchmark for aesthetic badcase discovery: it evaluates whether a model can jointly output an aesthetic score and a grounded flaw list organized by an 8-category taxonomy, with ground truth constructed by fully human annotation plus review. On the data side, we construct multi-category, multi-granularity, verifiable aesthetic badcase samples, each equipped with a problem description and localization of visible issues; on the model side, we train AesCase-8B with this recipe as a baseline, validating the usability of the benchmark and the effectiveness of the recipe.

Our contributions can be summarized as follows:

- We propose a score-guided five-tier construction method that simulates the human badcase distribution and balances annotation efficiency and quality, and build on it the AesCase-20k dataset (constructed via model pre-annotation plus human add/delete review) and the AesCase-bench benchmark (constructed via fully human annotation plus review).
- We train a baseline model AesCase-8B on mixed flaw-discovery and scoring data, jointly learning flaw discovery and score prediction, while maintaining competitive scoring performance on six IAA/IQA benchmarks, and generalizing badcases into repair suggestions via two-stage reasoning, without dedicated suggestion training data.
- This repair-oriented, interpretable feedback closes the loop from "where is the problem" to "how to fix it", making AesCase a plug-and-play model for image editing and aesthetic enhancement.

2 Related Work

Aesthetic scoring and captioning datasets. Conventional image aesthetic assessment datasets (Murray et al., 2012; Kong et al., 2016; He et al., 2022; Lu et al., 2014; He et al., 2023) are mainly built around overall scores or preference labels, and PARA (Ren et al., 2017) introduces personalization; later user-comment-supervised captioning datasets supplement textual explanations. However, both kinds of data lack structured descriptions of concrete visual problems: the former provides only scalar scores, the latter only overall impressions, making it hard for models to learn "where it is bad, why, and how to fix it".

Aesthetic MLLMs and critique. Multimodal large models have entered aesthetic understanding: Q-ReAlign (Q-Future, 2025) achieves scoring via discrete-level alignment, and AesExpert (Huang et al., 2024), ArtiMuse (Cao et al., 2026), VILA (Ke et al., 2023) combine scoring with fine-grained textual analysis, among which ArtiMuse already attends to critique. However, existing critique remains general commentary, lacking structured diagnosis centered on badcase + suggestion.

RL and reasoning-based IQA. Recent works (Aes-R1 (Liu et al., 2025), Q-Insight (Guo et al., 2025; Wu et al., 2025; Li et al., 2025b)) improve IQA via reinforcement learning (GRPO) and long chain-of-thought reasoning, at the cost of a complex pipeline and verbose, self-contradictory outputs. In contrast, AesCase directly learns the structured "score + 8-category flaw" output via pure SFT, without reward modeling or CoT.

Guidance, cropping, and editing benchmarks. Closer to AesCase, Venus (Hong et al., 2024; Du et al., 2026), although it can give suggestions, is limited to composition cropping and photography guidance, and cannot cover the broader range of aesthetic flaws (color, lighting, quality, depth, content, texture, realness). AesCase therefore constructs a general badcase taxonomy, structuring critique into a score, an 8-category flaw list, and integrated repair suggestions. Table 1 compares AesCase with existing works: most remain at image assessment, whereas AesCase directly drives aesthetic enhancement with structured badcase outputs.

Table 1: Comparison with existing aesthetic datasets/works. Fine-grained: ✓= fine-grained textual analysis, × = scalar score / labels only; Benchmark: whether a paired evaluation benchmark is provided. AesCase directly drives aesthetic enhancement with structured badcase outputs (8-category flaws + suggestions). The data-size column shows the sample scale of each dataset / work. For AesCase, 22K comprises 20k training images plus a 2K benchmark.

Dataset / Work	Data size	Fields	Fine-grained	Benchmark	Applications
AVA	255K	Score	×	✓	Image assessment
AADB	10K	Score + attributes	×	✓	Image assessment
TAD66K	66K	Score + attributes	×	✓	Image assessment
ArtiMuse	10K	Score + analysis	✓	✓	Image assessment
AVA-Captions	230K	analysis	✓	×	Image assessment
AesMMIT	22K	Score + analysis	✓	×	Image assessment
Venus	10K	Score + analysis + suggestion	✓	✓	Photography guidance
AesCase (ours)	22K	Score + flaw analysis + suggestion	✓	✓	Image assessment + aesthetic enhancement

3 AESCASE-BENCH

The core goal of AesCase is to enable models to recognize aesthetic badcases in real images. We define a badcase as an observable, explainable, local or global problem in an image that is related to aesthetic judgment or image-editing quality. A badcase can be a low-level quality flaw, such as blur, noise, compression artifacts, and poor texture; it can also be a high-level aesthetic flaw, such as compositional imbalance, unreasonable content, or insufficient realness. We use *badcase* and *flaw* interchangeably throughout the paper; badcase discovery therefore corresponds to flaw discovery.

As shown in Figure 2, we construct a taxonomy of 8 badcase categories to unify data construction, annotation, and evaluation. Given the lack of a unanimously agreed standard for aesthetic attributes (Jin et al., 2019; Ghosal et al., 2019), this taxonomy is induced from bad-image-pool annotation and finalized through expert discussion, yielding a balanced badcase distribution. The 8 categories are: composition, quality, lighting, color, depth, content, texture, and realness. Composition concerns layout and framing; quality concerns clarity, noise, and compression; lighting concerns exposure and light; color concerns color balance; depth concerns depth of field and focus; content concerns impact and interest; texture concerns surface texture and detail; and realness concerns naturalness and authenticity.

AesCase-bench is used to evaluate whether a model has truly learned to discover and explain aesthetic cases. Compared with benchmarks that only look at overall scores, AesCase-bench requires models to output structured answers covering the case type, the location of the problem, the cause analysis, and the suggestion direction. During evaluation, we report Coverage (the recall of ground-truth flaws), Precision (the fraction of model outputs that match ground-truth flaws), and F1 of flaw

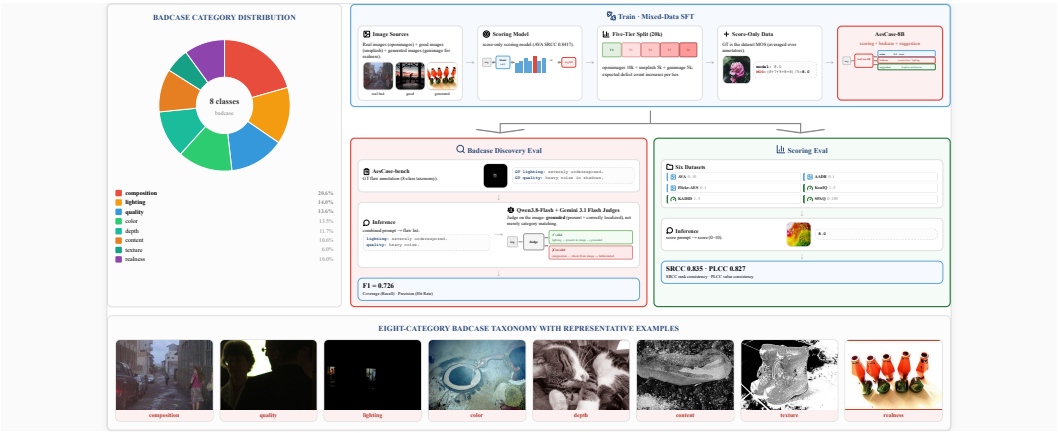

Figure 2: AesCase taxonomy and framework. Left: the 8-category badcase distribution (composition / lighting / quality dominate). Right: the framework—mixed-data training (openimages + unsplash + genimage, scored by score1dec and split into five tiers, SFT into AesCase-8B with scoring and badcase discovery) and two parallel evaluations (flaw-discovery F1 on AesCase-bench, SRCC/-PLCC on six datasets). Bottom: representative 8-category examples.

discovery, where an MLLM judge, in combination with the image, decides whether the model's flaws cover the ground truth, are real, and are accurately localized.

AesCase-bench contains 2000 images in total: 1000 low-score bad images (sampled from openimages + genimage by score1dec scores, with zero overlap with the training data) and 1000 high-score good images (sampled from Pexels, independent of the three training sources; see Table 5). The ground-truth flaws of the bad images are constructed by fully human annotation plus review: 3 annotators annotate flaws image-by-image and flaw-by-flaw (category + grounded description), with a flaw-level agreement of $\kappa = 0.8$, followed by cross-review and collective discussion to reach consensus, ensuring the ground truth is independent of any MLLM pre-annotation and cutting off the "evaluation GT and evaluation judge are from the same source" loop at its root.

4 Method

4.1 Score-Guided Data Construction

We adopt a two-stage score-guided data-construction strategy, where the upper bound on the number of flaws is given by:

$$k(\text{tier}) = \text{tier}, \quad \text{tier} \in \{0, 1, \ldots, 4\}. \tag{1}$$

- **Stage 1 (Score labeling):** a model trained on the aesthetic-scoring task (score1dec, SRCC 0.8417) uniformly predicts a 0–10 aesthetic score for each candidate image, and then images are split into 5 tiers by global score quantiles, where Tier 0 is the high-score good images (the top 20% by score, ≥ 8, from unsplash), and Tiers 1–4 are the low-score bad images evenly split from high to low score;
- **Stage 2 (Badcase labeling):** a per-tier prompt is designed to guide the model to find 0–x badcases (x increases with tier: Tiers 0 to 4 correspond to 0, 0–1, 0–2, 0–3, 0–4 flaws respectively). An MLLM generates the pre-annotations, outputting one "category: description" per line, or "none" when there are no flaws.

Labeling model. The scoring model (score1dec) used in the Score step is an independently trained score-only model (based on Qwen3-VL, one-decimal scoring, AVA SRCC 0.8417), serving only as a data-construction tool and being independent of the final AesCase-8B; AesCase-8B is trained via single-stage full SFT on the mixed five-tier data and score-only data, avoiding the catastrophic forgetting of staged training.

score1dec predicts scores with one decimal place; the granularity ablation on the number of predicted digits (integer vs. one decimal vs. two decimals) is reported in Appendix 7.1.

The annotation of each sample contains an aesthetic score and a flaw list (each item being a category plus a concrete description). For the data ratio, we uniformly adopt the three sources openimages, unsplash, and genimage at a 2:1:1 ratio to form 20000 training samples: openimages 10000 (general real images, covering the mid-low score segment), unsplash 5000 (high-score good images), and genimage 5000 (AI-generated images, covering realness/texture flaws). After mixing the three sources, we uniformly split by global score quantiles (Figure 5): Tier 0 (good images) is mainly contributed by unsplash high-score images, and Tiers 1–4 (bad images) are covered by openimages and genimage from high to low score. In addition, we perform image-level deduplication (perceptual-hash deduplication) among the three training sources, the six scoring evaluation datasets (AVA / AADB / Flickr-AES / KonIQ-10k / KADID-10k / SPAQ), and the 2000 AesCase-bench evaluation images (especially the openimages / genimage bad images that share sources with the training data), ensuring zero overlap between the training set and every evaluation set and ruling out data leakage.

Human add/delete review. Since LLM labeling has inherent noise (category misclassification, vague descriptions, score–flaw inconsistency), we add a human review step after automatic labeling: annotators review all 20000 samples image-by-image, deleting false-positive/nitpicky entries and supplementing missed flaws, with the Tier-0 "none" output human-confirmed. Every annotation also undergoes cross-review with iterative discussion to consensus. Review shows about 22% of LLM annotations are deleted and about 18% of final flaws are human-added; the human-judged flaw count is highly consistent with the LLM's $k(\mathrm{tier}) = \mathrm{tier}$ labeling (per-tier distribution overlap > 0.9), confirming the monotonic "lower score, more flaws" rule reflects human judgment rather than an injected format prior.

To strengthen realness discrimination, we introduce genimage AI-generated images to cover AI-artifact flaws that rarely appear in real images, and to recognize new artifacts introduced during multi-round editing. Mixing the three sources also balances the badcase categories: real-image sources (openimages / unsplash) mainly contribute composition / lighting / quality flaws, while genimage supplements texture / realness flaws, raising the realness share to about 10% and alleviating category long-tail.

4.2 Mixed-Data Supervised Fine-Tuning

The training data consists of two parts. The first part is the five-tier combined data, where the prompt of each sample requires the model to first output a 0–10 aesthetic score and then find 0–x badcases (x determined by the score tier); the second part is the score-only data, where the prompt only requires the model to output an aesthetic score kept to one decimal place. The two parts are mixed and trained with a unified language-modeling objective.

4.3 Suggestion Generalization

Note that the training supervision of AesCase-8B covers only flaw discovery and scoring, and does not explicitly include repair suggestions. The key insight is that aesthetic flaws are universal—flaw localization and attribution of "where it is bad and why" are shared across downstream tasks (image editing, photography guidance, content review), while only the output form of "how to fix it" varies. We therefore organize suggestion as two-stage reasoning: Stage 1 reuses badcase discovery to output a structured flaw list; Stage 2 takes this list as conditional input and, via a task prompt, guides the model to output task-specific suggestions (Figure 6), so suggestions are naturally grounded and require no extra training data. By modifying the system prompt, the same model serves different downstream tasks (editing instructions, photography guidance, content review) and can give either integrated suggestions or a local repair for a user-specified badcase, supporting interactive local editing.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Main experiment I: AesCase-8B trained on mixed data (20000 five-tier combined + score-only) is evaluated on AesCase-bench for badcase discovery, reporting flaw-discovery F1. **Main experiment II:** the same AesCase-8B is evaluated on six datasets (AVA (Murray et al., 2012), AADB (Kong et al., 2016), Flickr-AES (Lu et al., 2014), KonIQ-10k (Hosu et al., 2020), KADID-10k (Lin et al., 2019), SPAQ (Fang et al., 2020)) for scoring ability (SRCC / PLCC), verifying whether scoring ability is preserved.

Compared models: we compare general MLLMs (Qwen3-VL (Bai et al., 2025), GPT-5.5 (OpenAI, 2025), Gemini (Team, 2025)), traditional IAA/IQA methods (NIMA (Talebi & Milanfar, 2018), MUSIQ (Ke et al., 2021)), and aesthetic/quality models (ArtiMuse (Cao et al., 2026), Q-ReAlign (Q-Future, 2025), Q-Align (Wu et al., 2024), UniPercept (Hu et al., 2025), Aes-R1 (Liu et al., 2025), etc.).

Evaluation metrics: on AesCase-bench we report Coverage (recall of ground-truth flaws), Precision, and F1 of flaw discovery; on the scoring datasets we report SRCC and PLCC. All results use the same test split and inference settings. The flaw-discovery metrics are defined as follows:

$$\text{Coverage} = \frac{|C|}{|G|}, \quad \text{Precision} = \frac{|C|}{|M|}, \quad \text{F1} = \frac{2 \cdot \text{Coverage} \cdot \text{Precision}}{\text{Coverage} + \text{Precision}}, \tag{2}$$

where C is the set of ground-truth (GT) flaws covered by the model (outputs whose category matches GT and whose semantics are equivalent), G the full GT set, and M the total model outputs. F1 is aggregated over the 1000 bad images, and good images are reported separately by false-positive rate (Table 5); a bad image outputting "none" has Precision set to 1 by convention. The judge requires category consistency and semantic equivalence to count as coverage. We use the standard F1 as the main metric. In downstream editing, false positives drive unnecessary modifications, so precision fits repair-oriented feedback better than recall. To verify judge reliability, we human-re-check 200 sampled outputs: 3 annotators judge whether flaws cover GT and are real and localized, with inter-annotator $\kappa = 0.85$ and human–judge $\kappa = 0.82$, confirming the main F1 is human-verified.

Table 2: Main experiment I: flaw-discovery results on the 1000 bad images of AesCase-bench (dual-judge cross-validation). The main metric is the standard F1, with Coverage and Precision both computed with the GT-matching set $|C|$ as the numerator. AesCase-8B ranks first under both judges.

Model	Qwen3.8-Flash judge			Gemini 3.1 Flash judge		
	Cov ↑	Prec ↑	F1 ↑	Cov ↑	Prec ↑	F1 ↑
General MLLMs						
Qwen3-VL-8B	0.484	0.354	0.409	0.538	0.431	0.479
InternVL3-8B	0.256	0.314	0.282	0.363	0.481	0.413
Gemini 3 Pro	0.513	0.496	0.505	0.571	0.559	0.565
GPT-5.5	0.769	0.424	0.547	0.861	0.475	0.612
Aesthetic / Quality MLLMs						
AesExpert	0.337	0.376	0.355	0.590	0.423	0.492
Venus	0.300	0.255	0.276	0.480	0.305	0.373
ArtiMuse	0.546	0.296	0.384	0.681	0.350	0.463
UniPercept	0.462	0.383	0.419	0.586	0.466	0.519
Aes-R1	0.711	0.261	0.382	0.883	0.324	0.474
Q-ReAlign-Pro-9B	–	–	–	–	–	–
Ours						
AesCase-8B	0.729	0.724	0.726	0.835	0.823	0.829

Q-ReAlign-Pro-9B is a score-only model (outputting only discrete score levels) and does not output a structured flaw list, so its flaw-discovery entries (Table 2) are marked as "–".

5.2 MAIN FINDINGS

The results of Main experiment I show that AesCase-8B achieves a flaw-discovery F1 of 0.726 (Qwen3.8-Flash judge) on AesCase-bench, ranking first among all compared models (GPT-5.5 0.547, Gemini 3 Pro 0.505). The core conclusion is that AesCase-8B produces almost no false positives on high-score good images (0.02), indicating that the mixed data and the good-image "none" supervision indeed yield a non-nitpicking, grounded diagnostic ability whose output is consistent with the image content. Re-evaluating with the Gemini 3.1 Flash judge, AesCase-8B still ranks first (0.829), and the top-1 and top-4 rankings are consistent under the two judges, showing that the results do not depend on the choice of judge.

Further analyzing the failure modes of the compared models, we can divide them into two categories. The first is "piling-up" (Aes-R1, GPT-5.5, ArtiMuse): they trade a large number of flaws for coverage, but their precision is only 0.26–0.42, meaning over half of their outputs are redundant or hallucinated flaws, and they also pile up false positives on high-score good images (Table 5 shows GPT-5.5 averaging 4.05 flaws and ArtiMuse 4.17). The second is "miss-prone" (AesExpert, InternVL3-8B, Venus, UniPercept): their outputs are restrained but they miss flaws, with a recall of only 0.26–0.46; among them AesExpert, as a five-tier level model, barely outputs a structured flaw list. AesCase-8B is the only model that is strong at both ends: its recall of 0.73 and precision of 0.72 are both high, its F1 of 0.726 ranks first, and its good-image false-positive rate is only 0.02—it neither "over-reports imprecisely" like the piling-up models nor "under-reports incompletely" like the miss-prone models, but precisely, restrainedly, and groundedly localizes real flaws.

As corroborating evidence for scoring compatibility, the score-guided AesCase-8B still maintains competitive scoring on the six IAA/IQA datasets (Table 3), showing that the recipe does not sacrifice scoring alignment.

Table 3: Main experiment II: results on the six IAA and IQA datasets after mixed-data (five-tier combined + score-only) training. Each cell reports SRCC / PLCC.

Model	IAA			IQA		
	AVA	AADB	Flickr-AES	KonIQ-10k	KADID-10k	SPAQ
General MLLMs						
Qwen3-VL-8B	0.408/0.422	0.453/0.474	0.442/0.469	0.788/0.825	0.708/0.723	0.898/0.899
InternVL3-8B	0.405/0.351	0.472/0.362	0.511/0.491	0.741/0.809	0.657/0.636	0.872/0.868
Gemini 3 Pro	0.508/0.548	0.421/0.447	0.674/0.688	0.595/0.645	0.636/0.633	0.767/0.777
GPT-5.5	0.654/0.619	0.494/0.490	0.708/0.702	0.826/0.873	0.703/0.690	0.880/0.876
Traditional IAA / IQA						
NIMA	0.680/0.685	0.382/0.414	0.512/0.553	0.859/0.896	0.535/0.532	0.856/0.838
MUSIQ	0.816/0.818	0.354/0.373	0.539/0.572	0.937/0.948	0.585/0.602	0.855/0.854
Aesthetic / Quality MLLMs						
AesExpert	0.520/0.524	0.543/0.509	0.728/0.721	0.323/0.365	0.546/0.470	0.614/0.663
Venus	0.442/0.432	0.294/0.240	0.305/0.313	0.439/0.420	0.319/0.348	0.498/0.404
ArtiMuse	0.827/0.826	0.505/0.493	0.814/0.837	0.459/0.525	0.454/0.504	0.807/0.788
UniPercept	0.589/0.576	0.424/0.424	0.705/0.697	0.966/0.969	0.908/0.899	0.911/0.900
Aes-R1	0.662/0.670	0.542/0.539	0.697/0.724	–	–	–
Q-Insight	0.590/0.596	0.518/0.507	0.677/0.701	0.916/0.933	0.736/0.742	0.905/0.907
Q-Align	0.822/0.818	0.514/0.513	0.642/0.660	0.949/0.959	0.967/0.971	0.945/0.948
Q-ReAlign-Pro-9B	0.832/0.828	0.479/0.475	0.655/0.676	0.949/0.955	0.963/0.966	0.944/0.947
Ours						
AesCase-8B	0.835/0.827	0.716/0.737	0.805/0.815	0.948/0.957	0.951/0.955	0.928/0.932

Aes-R1 is not trained for IQA, so it cannot produce scores on KonIQ-10k / KADID-10k / SPAQ and the corresponding three columns are marked as "–"; its AVA / AADB / Flickr-AES columns are complete results.

The qualitative analysis (Appendix Figure 3) shows that AesCase-8B can point to specific regions, such as subject occlusion, background distraction, local blur, color shift, or compositional imbalance, rather than giving a vague "this image is not beautiful enough" comment; its flaw descriptions carry region-level, object-level, and cause-level information.

To quantify the contribution of badcase discovery to interpretability, we conduct a controlled experiment: on 100 bad images, we let the same model output three forms, namely "badcase + suggestion" (first localize concrete flaws, then give repair suggestions), "suggestion only" (only suggestions, no flaw localization), and "badcase only" (only flaw localization, no suggestions), and human voters blindly rate which form is the most interpretable and most preferred. The result is that badcase + suggestion wins with 91 votes, suggestion only gets 6 votes, and badcase only gets 3 votes (Figure 4). This shows that flaw localization and repair suggestions form an inseparable complete loop: first answer "where and why it is bad", then answer "how to fix it", which is the interpretable, actionable feedback.

5.3 Ablation Studies

We further use "trained on score-only data" (score-only) as a lower-bound baseline: this model is trained on a 30k sampled subset of the score-only data and only learns scoring, never seeing any flaw-discovery supervision; it reaches a scoring of 0.7945, but when asked to output flaws at evaluation time, its F1 is only 0.42, slightly above the zero-shot baseline (0.409) but far below the full AesCase. This shows that the five-tier combined data is necessary for flaw discovery, while the score-only data is necessary for scoring, and mixed-data training balances the two.

We verify the necessity of score-guidance tiering through ablations. On the one hand, under the "0–k" output, the model tends to output the upper bound k, so 0–k can be used to explicitly control that the number of flaws increases with tier; on the other hand, if we do not tier and only use a flaw-finding prompt (no score guidance), the model "nitpicks": it averages 3.31 flaws even on high-score good images and outputs *none* only 4% of the time; after tiering, Tier 0 outputs *none* (none ratio 100%) and Tiers 1–4 output 0–k flaws, thereby controlling the flaw-count distribution and avoiding nitpicking.

The above ablation proves the value of tiering only on the good-image side: it suppresses "nitpicking" on high-score good images, but it does not yet answer a more critical question: whether tiering contributes to the flaw-discovery F1 on bad images, or merely makes the model output less and thus appear "restrained" on the good-image side. To isolate this, we fix the base model and data scale (same Qwen3-VL-8B base, same 20k five-tier data), change only how the training data is composed, and train three variants: "un-tiered case-only" (no tiering), "tiered (Tiers 1–4 only)" (no Tier-0 good images), and "tiered (all five tiers)" (full AesCase).

As Table 4 shows, the three form a clear causal chain. The lowest score-only (trained on scoring data only) has an F1 of 0.42, only slightly above the zero-shot baseline (0.409) and far below the full model, showing that a model that only learned scoring and never saw flaw-discovery supervision has limited flaw-localization ability; flaw supervision is necessary, not optional. After removing tiering, case-only has an F1 of 0.589, but a good-image false-positive rate as high as 3.31, and it can find flaws but cannot tell good from bad images, nitpicking high-score images as well. After introducing tiering, F1 rises to 0.705 (+0.116), because the 0–k flaw-count constraint aligns the output distribution with the ground-truth flaw-count distribution, freeing the model from the freedom of "report as many as it wants". Finally, adding the Tier-0 good-image supervision drops the good-image false-positive rate from 0.83 to 0.02, while the bad-image F1 slightly rises to 0.726. This chain shows that tiering contributes "finding more precisely" (higher bad-image F1), while the Tier-0 good-image supervision contributes "not reporting what should not be reported" (lower good-image false positives).

Table 4: Ablation of score-guided tiering's contribution to bad-image F1 (same Qwen3-VL-8B base model, re-evaluated by the same Qwen3.8-Flash judge). The three case-discovery variants share the same 20k five-tier data; score-only is a lower-bound anchor on 30k sampled score-only data.

Training setting	flaw-discovery F1 ↑	Good-image false positives ↓
score-only (scoring data only)	0.42	0.33
un-tiered case-only (no tiering)	0.589	3.31
tiered (Tiers 1–4 only, no Tier-0)	0.705	0.83
tiered (all five tiers, full AesCase)	0.726	0.02

As corroborating evidence, we also conduct a small-scale human "delete" review (Appendix Table 7): the number of human-kept flaws decreases monotonically from 3.59 in the low-score tier to 0.21 in the high-score tier, showing that "the number of flaws decreases with score" matches human annotation intuition, and that using 0–k to control the flaw count in score guidance is not an arbitrarily imposed prior.

Finally, we push badcase discovery down to image-editing verification: we let each model output badcases, and uniformly use Gemini 3.1 Flash to generalize them into suggestions that drive GPT Image 2 editing (Zhang et al., 2023), evaluating along three dimensions: aesthetic score gain, edit fidelity, and human preference (Xu et al., 2023). The unified suggestion isolates the differences in each model's suggestion-generation ability, so this experiment measures the flaw-discovery quality of each model. We choose GPT Image 2 rather than open-source editing models (InstructPix2Pix (Brooks et al., 2023), Qwen-Image-Edit) as the editing executor, because the latter cannot faithfully follow fine-grained aesthetic repair instructions and cannot precisely execute region-specific repair actions such as "straighten the horizon" or "reduce the highlights". As Table 5 shows, AesCase-8B's edits achieve the best balance between aesthetic score gain (+2.4) and edit fidelity (LPIPS 0.2866), and receive the highest total votes (431 votes, 10 annotators $\times$ 100 images): the piling-up models (GPT-5.5, Aes-R1, ArtiMuse) have a higher Δ (+2.5$\sim$2.6) but heavily break fidelity (LPIPS 0.41$\sim$0.45), while the miss-prone models (AesExpert, Qwen3-VL-8B) preserve fidelity well (LPIPS 0.16$\sim$0.20) but barely repair (their Δ is only +0.7$\sim$1.5), showing that AesCase's grounded flaw localization yields repairs that are both precise and fidelity-preserving. The editing results of all models are shown in Appendix Figure 7 and Figure 8, and the AesCase-8B badcase-localization and repair-suggestion close-ups are shown in Appendix Figure 9.

Table 5: Downstream editing comparison: good-image false positives (↓), bad-image *none* ratio (%↓), aesthetic score gain (Δ ↑), edit fidelity (LPIPS ↓ / DINOv2 ↑), and total votes (↑, 10 annotators $\times$ 100 images).

Model	Good-image FP ↓	none %↓	Δ ↑	LPIPS ↓	DINOv2 ↑	Votes ↑
General MLLMs						
Qwen3-VL-8B	0.30	33.2	+1.5	0.2034	0.9283	57
InternVL3-8B	0.84	12.8	+1.9	0.3404	0.8939	38
Gemini 3 Pro	1.07	10.0	+2.0	0.2393	0.8942	69
GPT-5.5	4.05	0.0	+2.6	0.4511	0.8273	139
Aesthetic / Quality MLLMs						
AesExpert	0.29	66.3	+0.7	0.1567	0.9578	23
Venus	2.13	14.8	+2.0	0.3135	0.8899	43
ArtiMuse	4.17	3.9	+2.5	0.4066	0.8533	52
UniPercept	0.12	21.1	+2.1	0.2543	0.9049	79
Aes-R1	7.36	0.0	+2.6	0.4131	0.8637	69
Ours						
AesCase-8B	0.02	2.1	+2.4	0.2866	0.8935	431

6 CONCLUSION

This paper proposes AesCase, a research framework for aesthetic badcase discovery in images. We construct AesCase-bench and the accompanying training data, and train AesCase-8B. The experimental design has two parts: Main experiment I evaluates the badcase-discovery performance on AesCase-bench after training on AesCase data; Main experiment II evaluates the scoring performance on AVA, AADB, Flickr-AES, KonIQ-10k, KADID-10k, and SPAQ after training on mixed data (five-tier combined + score-only). By comparing with models such as Qwen3-VL and ArtiMuse, we validate the importance of badcase discovery for improving interpretable aesthetic diagnosis. AesCase aims to advance image aesthetic research from single scalar prediction to an interpretable, actionable visual-case discovery paradigm, whose flaw feedback can directly serve instruction-based image editing and aesthetic enhancement. In future work, AesCase's structured score and flaw feedback can serve as a reward model or critique model, providing fine-grained, interpretable training signals for reinforcement learning of image generation and editing models.

REFERENCES

Shuai Bai et al. Qwen3-vl technical report, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Shuo Cao, Nan Ma, Jiayang Li, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *CVPR*, 2026.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. In *CVPR*, 2026.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020.

Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. In *ICCV Workshops*, 2019.

Daya Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

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

Shuai He, Anlong Ming, Yaqi Li, et al. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *ICCV*, 2023.

James Hong, Lu Yuan, Michaël Gharbi, et al. Learning subject-aware cropping by outpainting professional photos. In *AAAI*, 2024.

Vlad Hosu, Hanhe Lin, Tamás Szirányi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 2020.

Xiwei Hu, Rui Wang, et al. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture, 2025.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, et al. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *ACM MM*, 2024.

Xin Jin et al. Aesthetic attributes assessment of images. In *ACM MM*, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *CVPR*, 2023.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

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, 2025a.

Weixia Li et al. Q-insight: Understanding image quality via visual reinforcement learning, 2025b.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *QoMEX*, 2019.

Boyang Liu, Yifan Hu, Senjie Jin, et al. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization, 2025.

Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM MM*, 2014.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

OpenAI. Gpt-5 system card, 2025.

Q-Future. Q-realign: Q-align-style framework with lightweight lmms, 2025. GitHub repository.

Jimmy Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *ICCV*, 2017.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 2018.

Gemini Team. Gemini 3: A family of highly capable multimodal models, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024.

Tianhe Wu et al. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank, 2025.

Jiazheng Xu et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Kai Zhang et al. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023.

7 Appendix

7.1 A.1 Implementation Details

A.1.1 Training Details

Hyperparameters. We use Qwen3-VL-8B-Instruct as the base model and adopt the official qwen-vl-finetune framework for single-stage SFT. Unlike staged training (pre-training followed by scoring fine-tuning), AesCase simultaneously learns score prediction and flaw discovery through one joint training: bad-image samples contain "score + flaw list", while good-image samples contain "high score + none". The training uses full SFT (the vision encoder, MLP, and LLM are all trainable, i.e., including vit); the specific pre-trained model and hyperparameters are shown in Table 6.

Table 6: Base model and training hyperparameters of AesCase-8B.

Configuration	Value
Base model	Qwen3-VL-8B-Instruct
Training method	single-stage full SFT (vision encoder + MLP + LLM, including vit)
Training data	20000 five-tier combined images + 306,795 score-only images (six scoring datasets)
Epochs	2
Learning rate	5e-6
Optimizer	AdamW (β_1=0.9, β_2=0.95, weight decay 0.1)
Distributed	DeepSpeed ZeRO-3, 8 $\times$ GPU

Score-only data. The score-only part of the training data consists of 306,795 images from six IAA/IQA datasets—AVA (235,560), Flickr-AES (36,499), SPAQ (10,125), AADB (8,458), KonIQ-10k (8,110), and KADID-10k (8,043)—each labeled with a 0–10 one-decimal aesthetic score. These score-only images are disjoint from the test splits used for scoring evaluation.

A.1.2 Score Prediction Granularity

Regarding the granularity of score prediction, we compare three prediction formats, namely integer, one-decimal, and two-decimal: integer prediction is too coarse to distinguish close scores; two-decimal labels are too fine, sparse, and unstable; one-decimal prediction strikes a balance between discriminability and annotation stability, achieving the highest AVA SRCC (0.8417). This choice is consistent with the digit-number ablation independently reported by RealQA (Li et al., 2025a): it finds that predicting one decimal place improves AVA SRCC by 0.094 over predicting the integer, while adding more decimal places yields only marginal gains; we further observe that two-decimal labels are sparse and unstable. Therefore score1dec uses one-decimal scoring.

7.2 A.2 Qualitative Comparison

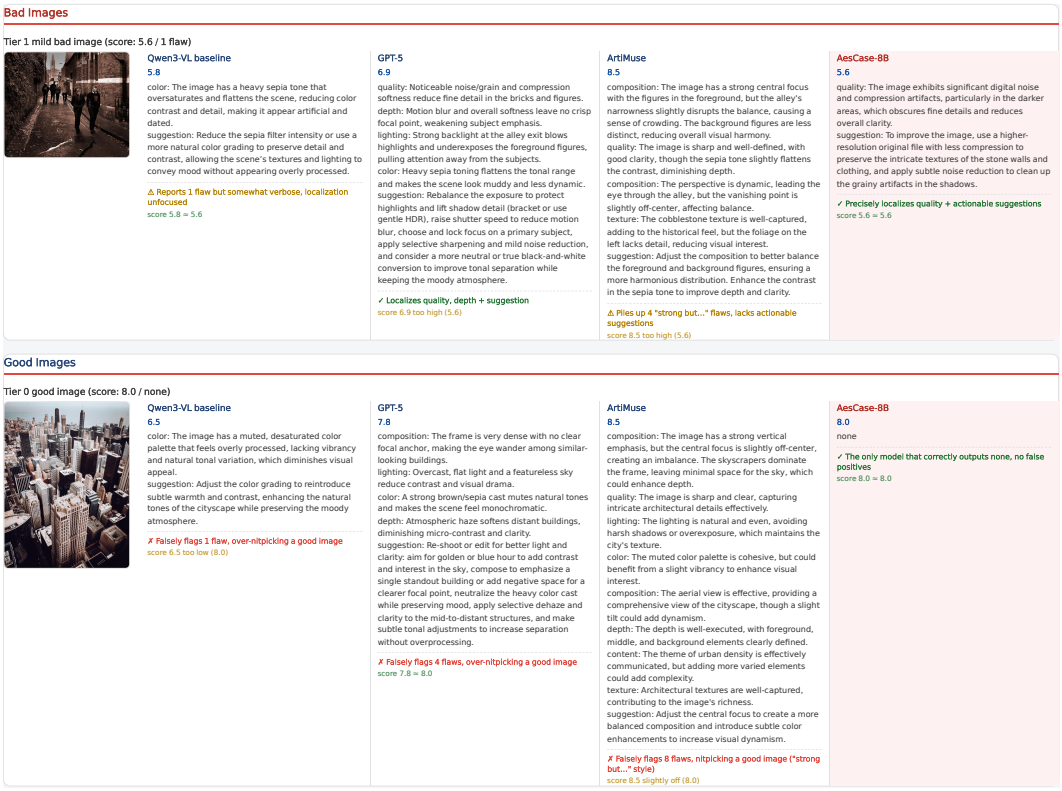

Figure 3: Qualitative flaw-discovery comparison among four models (Qwen3-VL baseline / GPT-5.5 / ArtiMuse / AesCase-8B) on Tier 1 bad + Tier 0 good images. AesCase-8B is the only model that outputs *none* on good images, precisely localizes flaws on bad images, and predicts scores close to score1dec.

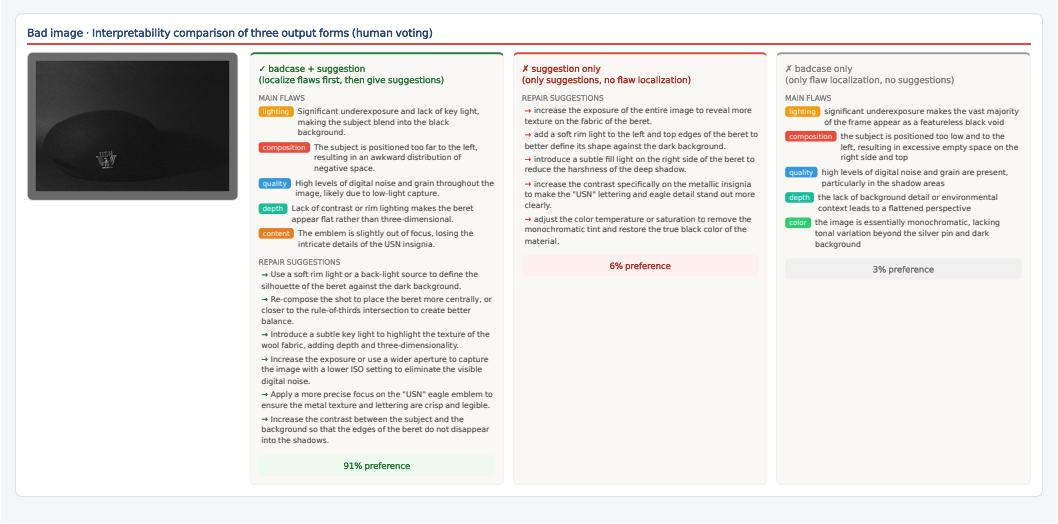

Figure 4: Interpretability comparison of three output forms: badcase+suggestion / suggestion only / badcase only. badcase+suggestion first localizes flaws (category + grounded description) then gives repair suggestions, forming a complete loop; suggestion only gives suggestions without flaws; badcase only localizes flaws without suggestions. Human voting shows badcase+suggestion wins by a large margin.

7.3 A.3 Human Deletion Audit

As corroborating evidence, we conduct a small-scale human "delete" review: gpt is asked to report a fixed 4 flaw candidates for each of 200 images in each of the five score tiers, and human annotators only delete (without adding). The number of human-kept flaws decreases monotonically from 3.59 in the low-score tier to 0.21 in the high-score tier, where the flaws reported by the model on bad-image tiers are almost all kept (real), while those on the good-image tier are almost all deleted (nitpicking). This further shows that "the number of flaws decreases with score" matches human annotation intuition, and that using 0–k to control the flaw count in score guidance is not an arbitrarily imposed prior.

Table 7: Under the human "delete" review, the number of kept flaws decreases monotonically with score tier (gpt reports a fixed 4 candidates per image, 5 tiers $\times$ 200 images, 1000 images in total).

Tier	Human avg. kept flaws	Retention rate
Tier 0	0.21	5.3%
Tier 1	1.03	25.8%
Tier 2	2.17	54.3%
Tier 3	3.19	79.8%
Tier 4	3.59	89.8%

7.4 A.4 Prompts

This section gives the key prompts used in evaluation and data construction.

Badcase discovery prompt (Table 2 flaw-discovery evaluation, unified across all models)

```
You are an image aesthetics assessment model. Find the main aesthetic
    flaws in this image.
For each flaw, assign a category and write a specific grounded
    description.
```

```
Categories:
- composition: layout and framing
- quality: image quality
- lighting: exposure and light
- color: color balance
- depth: depth of field and focus
- content: impact and interest
- texture: surface texture and detail
- realness: naturalness and authenticity

Descriptions must be specific and grounded: point to specific elements,
    locations, and severity.
List only real, visible, significant flaws. Do not nitpick or fabricate
    flaws.

Output one flaw per line, format "category: description". If no flaws,
    output only "none".
```

Scoring prompt (Table 3 six-dataset scoring)

```
Aesthetics: You are an image aesthetics assessment model. Rate the
    overall aesthetic quality of this image on a scale from 0 to 10.
    Return only a single number with one decimal place.

Quality: You are an image quality assessment model. Rate the overall
    perceptual quality of this image on a scale from 0 to 10. Return only
     a single number with one decimal place.
```

Judge prompt (the LLM judge for badcase F1, matching model output to GT flaws)

```
Above is an image. Below is GT (ground-truth flaws) and a Model's output
    describing aesthetic flaws in this image.

GT (ground truth):
0. composition: ...
1. lighting: ...

Model output:
0. composition: ...
1. quality: ...

Match each Model output flaw to the GT flaw it corresponds to (same flaw,
     correctly located).
Output one line per Model output flaw, in order: "<model_index>: <
    gt_index>", where gt_index is the matched GT index.
If a Model output matches no GT (redundant, hallucinated, or a real but
    unlisted flaw), output "<model_index>: none".
Each GT can be matched by at most one Model output.

Output only the mapping lines:
0: 0
1: none
```

Suggestion prompt (unified suggestion generation, stage 2 of two-stage reasoning)

```
You are an image aesthetics assessment model. Here are the aesthetic
    flaws identified in this image:
{flaws}

If the flaws above are 'none' (no flaws found), output 'suggestion: none
    '.
```

```
Otherwise, give concrete repair ACTIONS for these flaws as a short
    paragraph starting with 'suggestion: '.
Do NOT restate or describe the flaws. Be specific and grounded.
```

7.5 A.5 Score Distribution

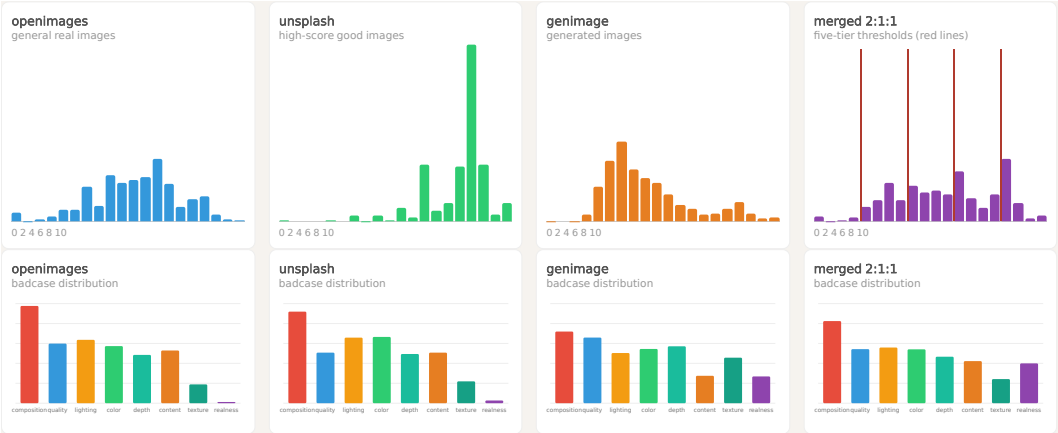

Figure 5: The score1dec score distributions of the three datasets (openimages / unsplash / genimage) and after 2:1:1 merging. The three sources respectively cover the low (genimage), mid-low (openimages), and high (unsplash) tiers; after merging, images are split into five tiers by global score quantiles (red lines).

7.6 A.6 Suggestion Generalization

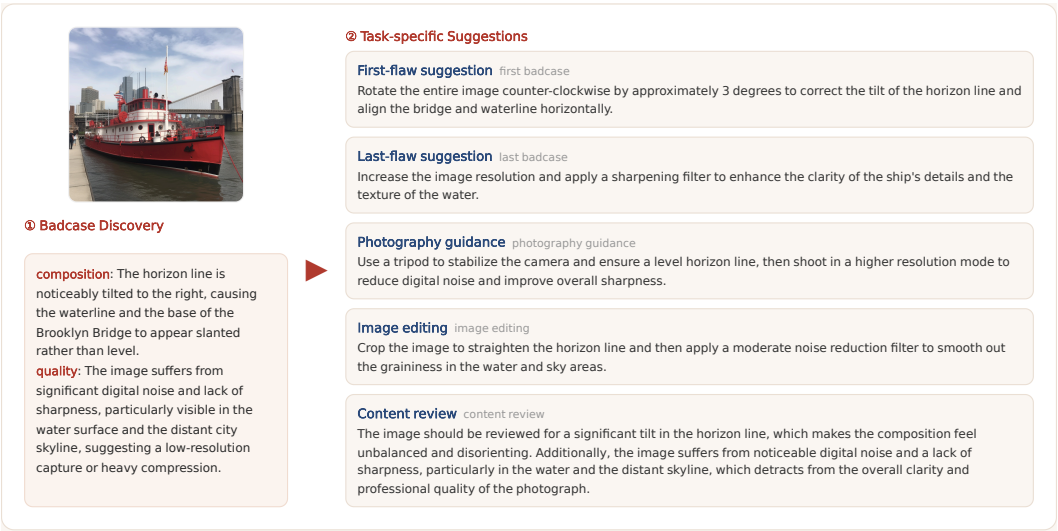

Figure 6: Suggestion generalization example. Given an image, the model first performs badcase discovery (Stage 1, localizing "where it is bad and why"), then modifies the prompt toward different downstream tasks (Stage 2): local suggestions (First-flaw / Last-flaw), photography guidance, image editing, and content review. The same flaw discovery generalizes to multiple tasks without per-task suggestion data.

7.7 A.7 EDITING RESULTS

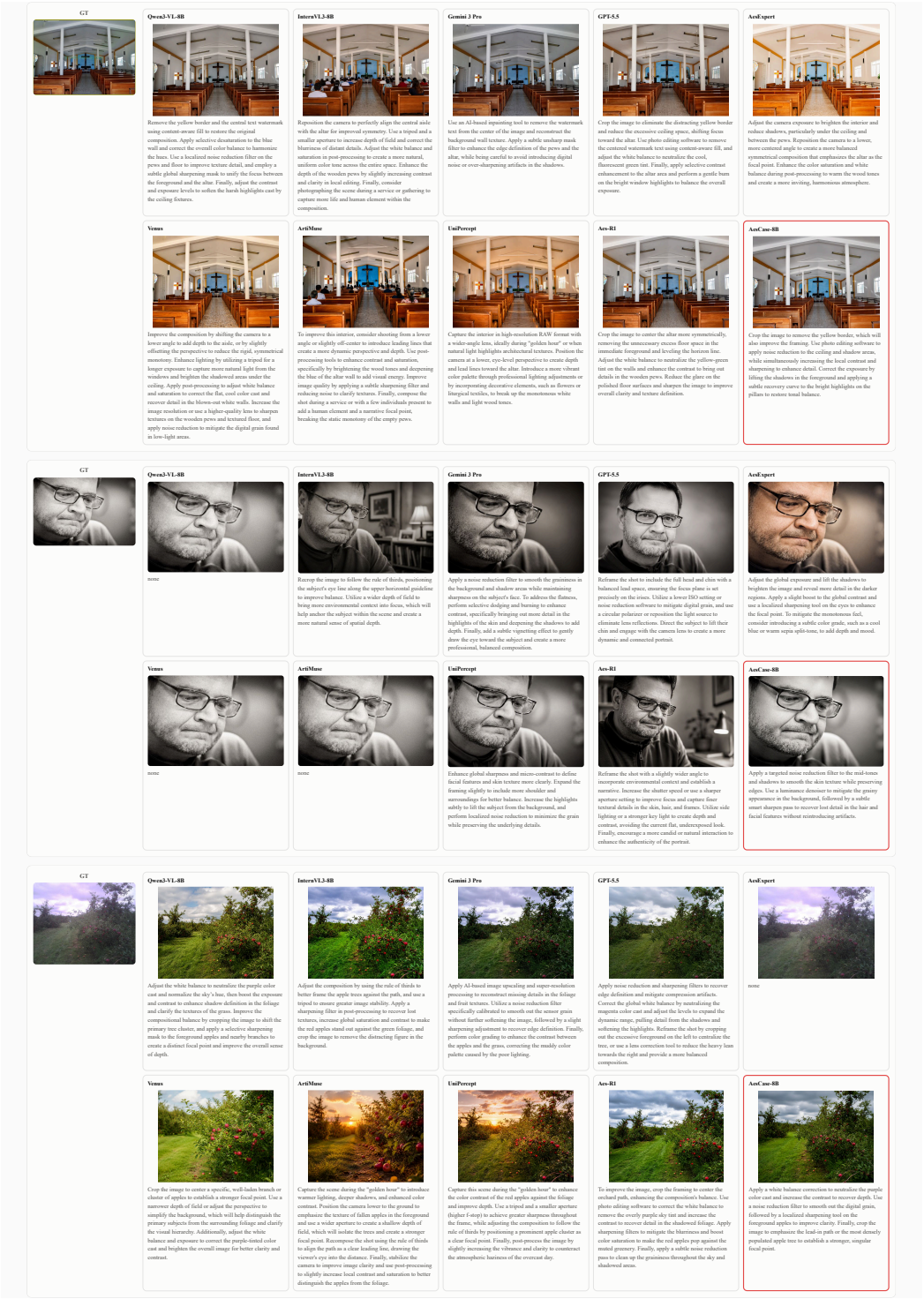

Figure 7: Image editing results (Fig A, 4 AesCase-bench bad images). Left is the GT original, right are the editing results of each model (uniform Gemini 3.1 Flash suggestion driving GPT Image 2 editing), with AesCase-8B in the red box.

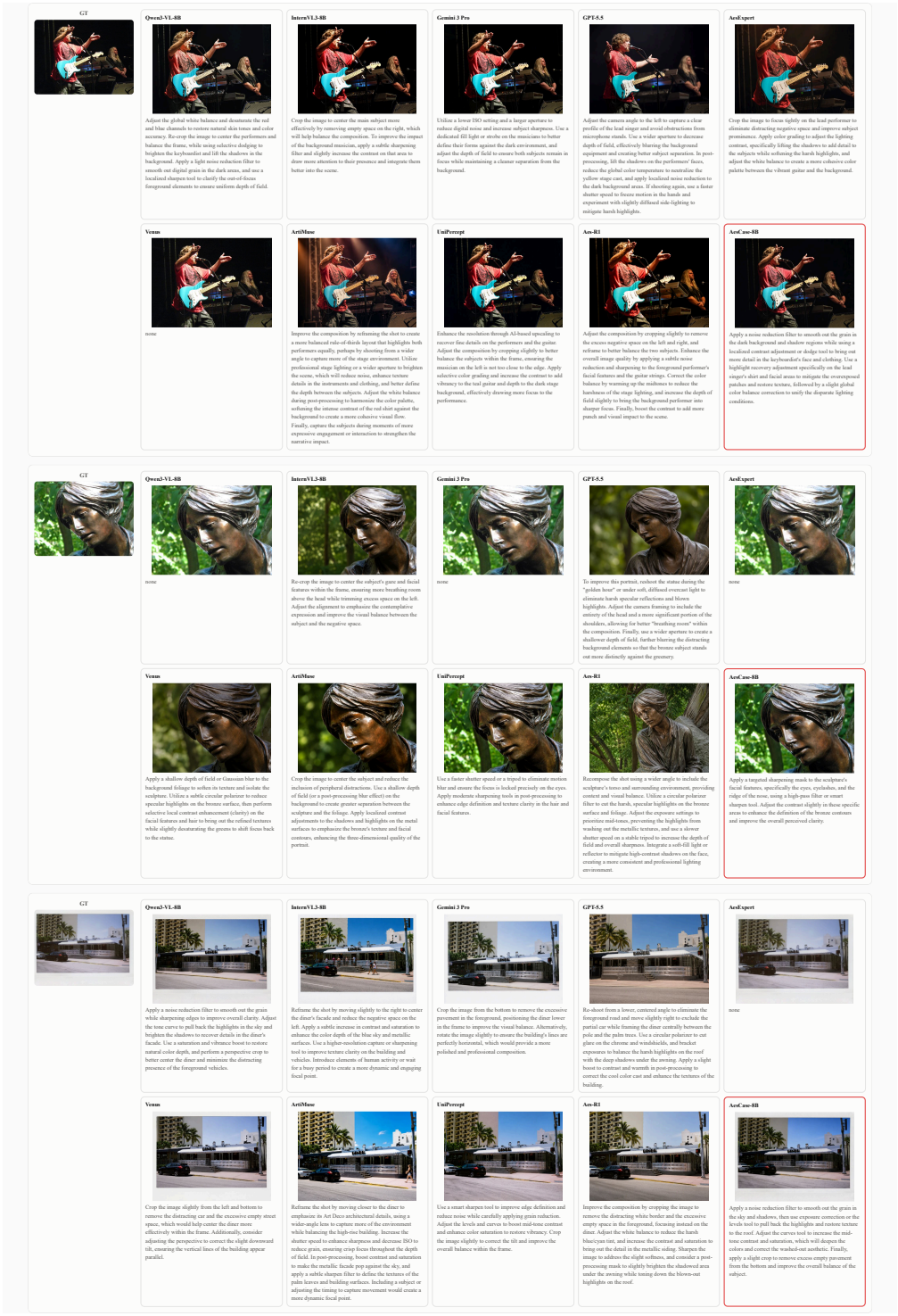

Figure 8: Image editing results (Fig B, 4 AesCase-bench bad images). Left is the GT original, right are the editing results of each model (uniform Gemini 3.1 Flash suggestion driving GPT Image 2 editing), with AesCase-8B in the red box.

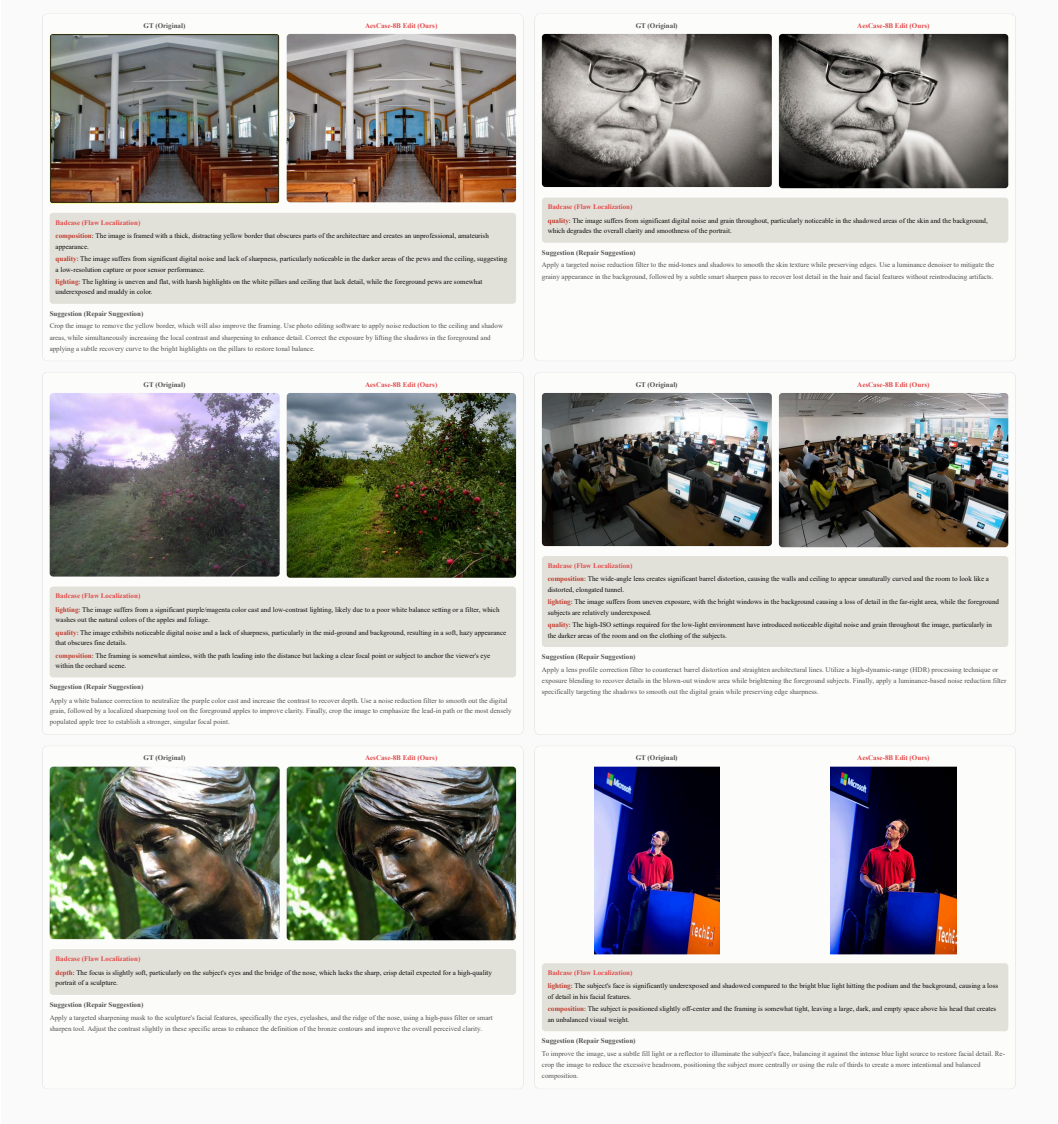

Figure 9: AesCase-8B editing close-ups (6 AesCase-bench bad images). Left is the GT original, right is the AesCase-8B edit (red box), with AesCase's badcase flaw localization (category + grounded description) and repair suggestion below, covering flaw types such as composition, quality, lighting, color, and depth.

7.8 A.8 Use of AI Assistant

We used an AI assistant during the writing of this paper, solely for language polishing and text refinement. It was not involved in any research content, experimental design, data analysis, or conclusion generation.

Reproducibility Statement

To promote reproducibility and future research, we will open-source the following resources: the AesCase-8B model weights, the AesCase-bench evaluation set, and the complete data-construction and evaluation code (including the score-guided five-tier construction, labeling, and judge scripts), all released under an open-source license.