AESTHETIC VISUAL REASONING WITH MULTIPLE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Aesthetic evaluation in photography intrinsically depends on collections rather than isolated images, as a photographer's style and intent emerge through recurring patterns across multiple works. While recent Multimodal Large Language Models (MLLMs) excel at single-image or pairwise aesthetic assessment, they struggle with group-level reasoning due to a lack of annotated datasets and fundamental limitations in handling long visual contexts. To enable the systematic study of this problem, we introduce VisGroup, the first large-scale dataset for group-level aesthetic visual reasoning, comprising approximately 1.4 million diverse image groups, along with a comprehensive benchmark (VisGroup-Bench). Through systematic analysis, we uncover unique properties of MLLMs in multi-image aesthetic contexts, notably a pronounced positional sensitivity. Building on these insights, we fine-tune a dedicated multi-image MLLM and propose a lightweight, model-agnostic visual triager that re-orders images based on predicted difficulty. This test-time optimization mitigates position bias and improves reasoning on challenging images without requiring architectural changes. Extensive experiments demonstrate that our approach significantly outperforms existing models, establishing a robust foundation and a strong baseline for future research in group-level aesthetic evaluation. ***We will release the model and dataset.***

1 INTRODUCTION

In professional photographic workflows, aesthetic evaluation rarely happens in isolation. Photographers routinely select the best works from hundreds of candidates, compare subtle variations across similar shots, or curate visually coherent portfolios. These tasks demand aesthetic reasoning across multi-image collections. However, while recent advances in Multimodal Large Language Models (MLLMs) (Liu et al., 2023; OpenAI, 2023; Bai et al., 2025) have driven significant progress in single-image critiques (Wu et al., 2024a; Huang et al., 2024a; Wu et al., 2024b; Qi et al., 2025) and pairwise comparisons (Wu et al., 2024c), group-level aesthetic reasoning remains a largely unexplored frontier.

Transitioning from single-image to group-level aesthetic understanding poses substantial challenges, primarily due to two bottlenecks. *First*, the multi-image reasoning capacity of current open-source MLLMs is fundamentally constrained. While open source MLLMs like Qwen2.5-VL (Bai et al., 2025) handle two to four images reasonably well (Wang et al., 2024), their performance degrades rapidly as the visual context expands to dozens of photos. This limitation is particularly detrimental to aesthetic evaluation, which demands careful visual inspection of fine-grained, low-level attributes, such as exposure, focus, and color grading, that easily get lost in long visual contexts. *Second*, the community lacks high-quality group-level datasets. Existing aesthetic datasets (Wu et al., 2024b; Qi et al., 2025; Huang et al., 2024a) primarily focus on individual images. Scaling this manual paradigm to large groups requires prohibitive amounts of time and human resources, creating a severe data drought for this task.

To bridge this gap, we introduce the task of group-level aesthetic visual reasoning and present *VisGroup*, the first large-scale dataset for this task. VisGroup comprises approximately **1.4M** high-quality and diverse image groups. To overcome the manual annotation bottleneck, we source our data from the Digital Photography Challenge (DP Challenges[1]). While prior work (Qi et al., 2025)

[1] https://www.dpchallenge.com/

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.

used this platform's user comments for single-image critiques, we uniquely leverage its rich, overlooked metadata, including author identity, photographic themes, and creative objectives. By designing a refined pipeline that extracts knowledge from individual critiques and clusters images by aesthetics-related topics (*e.g.*, style, technique, storytelling), we naturally construct diverse, real-world image collections. To rigorously evaluate model capabilities across a broad spectrum of photographic knowledge, we also introduce a comprehensive benchmark, *VisGroup-Bench*.

Leveraging VisGroup, we fine-tune a dedicated MLLM for **A**esthetic **VIS**ual reasoning at the group level, which we call **AVIS**. Rather than focusing on heavy architectural modifications, our primary contribution is establishing the data infrastructure and task formulation necessary to systematically study this problem. Through systematic analysis of AVIS and other MLLMs, we uncover unique behaviors in multi-image aesthetic contexts, most notably, a severe sensitivity to image positioning. Motivated by these findings, we introduce the Visual Triager, a lightweight, model-agnostic test-time optimization. By dynamically reordering input images based on their perceived difficulty, this strategy effectively mitigates positional bias and significantly enhances reasoning on challenging multi-image inputs without requiring any changes to the underlying model architecture.

We summarize our contributions as follows:

1. We introduce the task of group-level aesthetic visual reasoning and release *VisGroup*, a dataset of **1.4M** image collections. These groups are curated by extracting structural metadata (*e.g.*, themes, author intent) from photography forums to accurately reflect real-world curation.

2. We establish *VisGroup-Bench* to rigorously evaluate multi-image reasoning across key aesthetic dimensions, such as photographic style, technique, and storytelling, providing a standardized baseline for the community.

3. Our systematic analysis reveals that MLLMs suffer from positional bias in long aesthetic contexts. To address this, we propose a lightweight test-time optimization that dynamically reorders input images based on predicted difficulty, effectively mitigating positional bias without requiring architectural modifications.

2 Related Work

Aesthetic Visual Reasoning with MLLMs. Recent advances in MLLMs have elevated aesthetic visual understanding from traditional regression-based image assessment (IAA) (Ren et al., 2017; Murray et al., 2012; Yang et al., 2022) to open-ended tasks, including aesthetic conversation (Wu et al., 2024b; Huang et al., 2024a; Zhou et al., 2024; Huang et al., 2024b) and detailed photo critiques (Qi et al., 2025). Despite this progress, these approaches remain confined to single-image contexts, neglecting the curation and group-level reasoning critical to real-world photographic workflows. While recent efforts like Co-Instruct (Wu et al., 2024c) introduce comparative settings, they are limited to pairwise comparisons of a few (typically two to four) images. In contrast, our work introduces group-level aesthetic reasoning, requiring models to synthesize aesthetic patterns, themes, and styles across tens or even hundreds of images simultaneously.

Multi-image Understanding with MLLMs. Recent research on multi-image reasoning in MLLMs (Zong et al., 2024; Lin et al., 2023; Ye et al., 2023; You et al., 2023; Chen et al., 2023b; Peng et al., 2023; Tong et al., 2024; Chen et al., 2023a;c; Bai et al., 2023a; Dai et al., 2023) has predominantly focused on developing benchmarks (Liu et al., 2024a; Wang et al., 2024; Meng et al., 2024; Zhao et al., 2024; Cheng et al., 2025) to evaluate inter-image relationships, such as visual correspondence or semantic coherence. While state-of-the-art MLLMs (Bai et al., 2025) perform competitively on these tasks, the evaluation is typically constrained to short visual contexts of two to four images. Group-level aesthetic reasoning fundamentally departs from this paradigm. It demands not only the discovery of subtle aesthetic patterns within a massive set but also the abstraction of higher-level stylistic relationships across different photo groups. Furthermore, as the visual context expands to dozens of images, existing MLLMs suffer severe degradation in reasoning capacity. While recent studies (Tian et al., 2025) have identified position bias in short-context semantic tasks, our work exposes these vulnerabilities in long-context aesthetic reasoning and mitigates them through a novel, model-agnostic test-time optimization. We provide a detailed discussion of furhter related work in Appendix A.

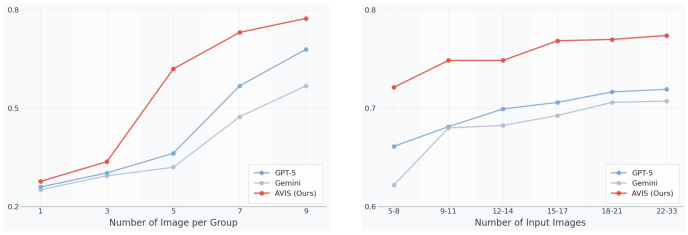


Figure 1: **Photograph style emerges in groups.** The y-axis displays the zero-shot accuracy of leading MLLMs (GPT-5, Gemini) alongside our proposed AVIS model. **Left:** The accuracy of aesthetic style prediction increases consistently as the number of images in a group expands. **Right:** Authorship identification accuracy improves with larger image groups, demonstrating that aesthetic evaluation scales positively with broader visual context.

3 VisGroup: A Dataset for Group-Level Aesthetic Visual Understanding

From professional portfolio curation to selecting images for social media, aesthetic evaluation is inherently a group-level task. A photographer's style, intent, and visual language often become apparent only through a collection of works.

To validate this intuition, we conducted two analytical experiments using MLLMs. In the first (Multi-Group Reasoning), the model is tasked with matching a source group of n images to one of four candidate groups captured by the same photographer (Fig. 4, bottom right). As shown in Fig. 1 (left), matching accuracy increases consistently as the group size expands. In the second experiment (Single-Group Critique), the model selects the option that best summarizes a group's stylistic characteristics (Fig. 4, bottom left). Fig. 1 (right) demonstrates that stylistic identification accuracy scales positively with the number of images. These findings confirm that robust group-level aesthetic reasoning is crucial for advanced photographic applications, motivating the creation of VisGroup.

3.1 Discovering Photo Groups

A major challenge to advancing group-level aesthetic visual understanding is the severe scarcity of training data that includes both image groups and rich aesthetic annotations. Existing datasets typically fall into two restricted paradigms: they either provide abundant single-image critiques without any group-level organization (Zhong et al., 2023; Qi et al., 2025; Wu et al., 2024b), or they offer grouped collections that lack detailed aesthetic descriptions, such as Unsplash[2].

Acquiring detailed aesthetic annotations from scratch is inherently costly, as it relies heavily on extensive human labeling (Wu et al., 2024b) or scraping massive volumes of user comments (Zhong et al., 2023; Qi et al., 2025). Consequently, retrospectively annotating large-scale collections like Unsplash is financially and logistically infeasible. However, discovering latent grouping structures within datasets that *already* contain abundant aesthetic annotations offers a highly cost-effective and scalable solution.

Guided by this insight, we revisit existing aesthetic datasets and specifically target the DP Challenge platform. Crucially, DP Challenge provides not only extensive user comments but also rich structural metadata, such as photographer identity, gallery category, and competition theme, as illustrated in Fig. 2 (left). These metadata attributes naturally organize the photos into semantically meaningful baseline groups. A brief statistical overview of these initial groupings is provided in Appendix C (Table 4).

Specifically, *Author* refers to the photographer of the image. A *Challenge* denotes the thematic competition to which the photo was submitted (e.g., *Sunset or Sunrise*, shown in Fig. 2, left). *Gallery* represents the photo's broader category (e.g., landscape or portrait), while *galleries* denote combinations of these categories (e.g., an image belonging to both *Landscape* and *Nature*, Fig. 2, left).

[2] https://unsplash.com/data

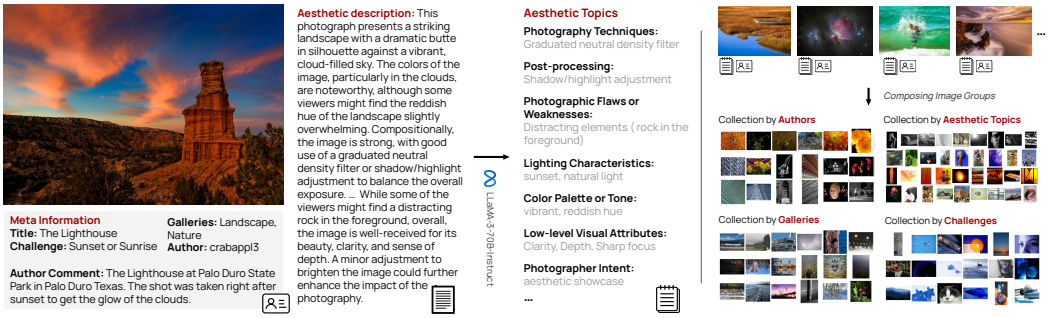


Figure 2: **Semantic Clustering via Key Phrase Extraction. Left:** Each image is initially associated with structural metadata (*e.g.*, challenge themes, galleries) and a unified photo critique derived from raw user comments. **Right:** An LLM processes these text critiques to extract discrete key phrases categorized into 17 predefined aesthetic topics. These semantic tags are then merged with the structural metadata to organize isolated images into fine-grained, aesthetically cohesive groups.

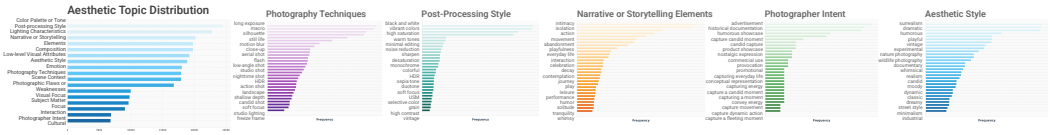


Figure 3: **Distribution of Aesthetic Topics.** Distribution of the 17 aesthetic topics across the dataset, alongside key phrase distributions for five representative topics.

Although this metadata provides a natural baseline for grouping images, it is insufficient for constructing high-quality, fine-grained aesthetic collections due to two primary limitations:

(1) Coarse-Grained Categorization. Challenge and gallery labels are often too broad. Categories such as *portrait* encompass highly diverse semantic and visual content, varying dramatically by scene (e.g., family snapshots vs. travel photography) or by composition (e.g., tight close-ups vs. full-body shots). Such internal heterogeneity makes the labels inadequate for cohesive stylistic grouping.

(2) Limited Cross-Topic Diversity. Relying solely on broad categories not only weakens intra-group coherence but also artificially restricts cross-topic diversity. For instance, when sampling solely by gallery labels, portraits and landscapes will rarely appear in the same group. However, despite depicting entirely different subjects, these images might convey identical emotional tones or share highly consistent lighting styles.

Discovering Groups from Photo Critiques. Beyond structural metadata, the aesthetic descriptions associated with each photo offer highly nuanced, fine-grained information. These texts span a wide spectrum of visual knowledge, ranging from low-level technical details to high-level emotional narratives, making them an ideal signal for fine-grained grouping.

We extract this signal through a systematic, LLM-driven pipeline. We begin with the raw DP Challenge data (Zhong et al., 2023), which contains multiple unprocessed user comments for each image. Following the approach of PhotoEye (Qi et al., 2025), we first employ an LLM to consolidate these disjointed comments into a single, unified photo critique, as shown in Fig. 2 (left).

Because directly clustering hundreds of thousands of long-form, free-text critiques is noisy and unreliable, we introduce an intermediate abstraction step. We prompt an LLM to parse each unified aesthetic description into a set of discrete key phrases. Each generated key phrase is strictly categorized into one of 17 predefined aesthetic topics, carefully curated to capture the most significant dimensions of aesthetic evaluation. The overall distribution of these 17 topics, along with detailed key phrase distributions for five sampled topics, is presented in Fig. 3.

Images sharing identical key phrases are then clustered together, yielding highly diverse groups that accurately reflect complex aesthetic relationships. To ensure a cohesive final dataset, we seamlessly integrate the initial structural metadata (challenges and galleries) into this new key-phrase frame-

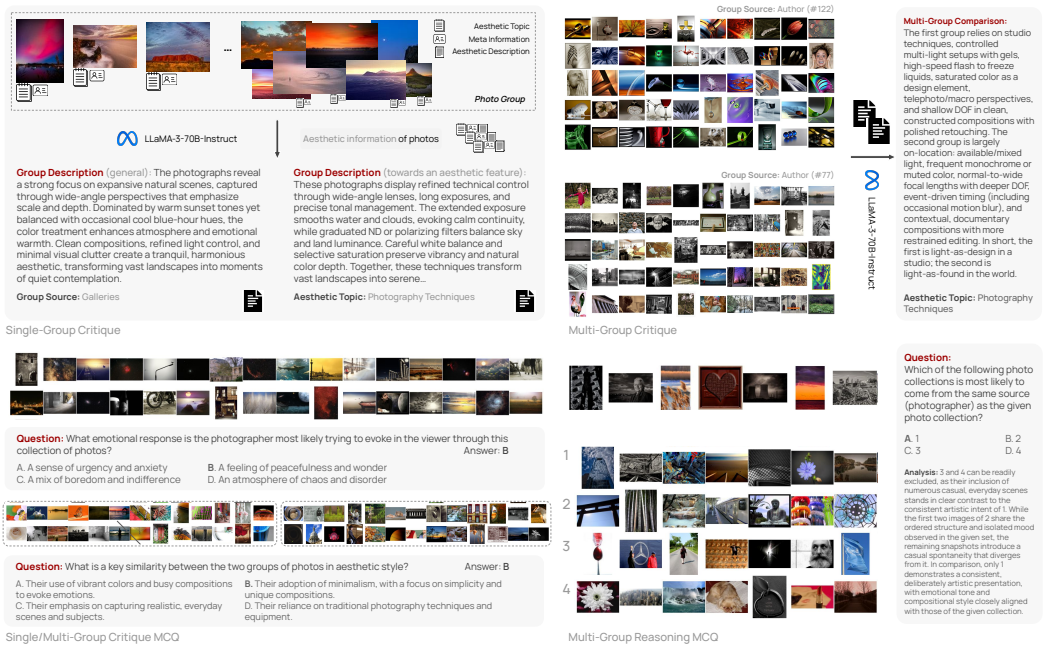


Figure 4: **Generating Multi-Level Aesthetic Annotations. Top Left:** Synthesizing a comprehensive single-group critique from aggregated metadata and individual descriptions. **Top Right:** Constructing multi-group comparative critiques by pairing single-group data. **Bottom Left:** Examples of generated Critique MCQs summarizing group features. **Bottom Right:** A Group-Reasoning MCQ requiring the model to match a candidate group to a source based on shared authorship or style. *Note: The textual analysis in the bottom-right example is provided for illustration only and is not included in the actual dataset.*

work. For example, the gallery label *nature* is subsumed as a key phrase under the *Scene Context* topic, as illustrated in Fig. 2. Ultimately, this pipeline successfully transforms disparate single-image critiques and broad metadata into a highly structured collection of aesthetically unified image groups.

3.2 Generating Multi-Level Aesthetic Annotations

Having structured the image collections, we next derive diverse, multi-level annotations to facilitate MLLM instruction tuning. The overall procedure is illustrated in Fig. 4. Because the base collections can contain thousands of images, we construct our dataset's basic units by sampling batches of 10–50 images without replacement for each key phrase. From these units, we generate four distinct types of data:

Single-Group Critiques. For each sampled photo group, we aggregate its metadata, aesthetic descriptions, and assigned topics. We then prompt LLMs to carefully review this text and synthesize a comprehensive critique characterizing the group's unifying features. To enrich the diversity of QA formats, we also prompt the LLM to provide topic-specific analyses (e.g., focusing strictly on photographic technique, as shown in Fig. 4, left) and concise, short QA pairs (e.g., *"These photos share a common characteristic of a soft gradient sky tint"*).

Multi-Group Comparisons. Real-world photographic workflows often require comparing stylistic differences across photographers or tracking stylistic evolution. To support this, we construct comparative data by sampling pairs from our single-group descriptions (Fig. 4, top right). For author-based groups, we randomly pair works from two distinct photographers. Crucially, for topic-based groups, we enforce a strict alignment constraint: both groups must originate from the same broad aesthetic topic. This ensures the comparisons are highly informative (e.g., contrasting "natural lighting" versus "studio lighting"). Without this constraint, the model would be forced to compare semantically irrelevant aspects, such as still-life lighting techniques against street photography emotions, yielding uninformative and misaligned data.

Aesthetic Multiple-Choice Questions (MCQs). Based on the generated descriptions, we construct two types of MCQs to explicitly test reasoning capabilities: (1) *Group-Critique MCQs*, which require the model to identify the correct stylistic summary for a group or group pair; and (2) *Group-Reasoning MCQs*, where the model is presented with a source group and four candidate groups, and must identify the candidate created by the same photographer or sharing the exact aesthetic trait. Examples are shown in Fig. 4.

Auxiliary Data for Foundational Understanding. To support fundamental multi-image capabilities, we also construct three types of simpler auxiliary data:

Single-Photo Judgment: A binary QA task determining if a photo satisfies a specific aesthetic key phrase. **Pair-Wise Comparison:** A two-photo QA task contrasting opposite key phrases within the same topic (e.g., under *Lighting Characteristics*: "Natural Lighting" vs. "Rembrandt Lighting"), asking the model to select the correct index. **Multi-Photo Referring:** An extension to multi-photo contexts where the model must identify 1–15 positive photos matching a target key phrase from a mixed pool of mutually exclusive negative samples. These auxiliary tasks build basic multi-image grounding and fundamental aesthetic perception (further details in Appendix B). The integration of all data types constitutes the complete **VisGroup** dataset, with detailed statistics provided in Appendix C (Tab. 6).

3.3 VISGROUP-BENCH

To evaluate the group-level aesthetic reasoning capabilities of MLLMs, we construct a comprehensive benchmark, **VisGroup-Bench**, comprising 1,291 high-quality Single-Group and Multi-Group multiple-choice questions (MCQs).

Derived from the highly diverse DP Challenge platform, the benchmark inherently encompasses thousands of photographic themes across all major genres, featuring images contributed by hundreds of thousands of users. This vast diversity naturally mitigates overfitting to specific data distributions. Following the pipeline described in the previous sections, we generate MCQs from individual photo critiques and metadata, which are then subjected to rigorous quality filtering. A standard train-test split based solely on unseen images is insufficient to demonstrate robust aesthetic comprehension. To ensure VisGroup-Bench evaluates true zero-shot reasoning, the benchmark is strictly disjoint from the training set at a structural level, with *no* overlap in authors or aesthetic features between the training and test sets. This strict separation allows us to test whether it can abstract its learned aesthetic knowledge and successfully extend it to novel stylistic concepts and unfamiliar photographic intents. All benchmark data have been manually checked. More details regarding the benchmark construction are available in Appendix C.

4 METHOD: AVIS FOR AESTHETIC VISUAL REASONING

Building upon the VisGroup dataset, we introduce **AVIS** (**A**esthetic **VIS**ual reasoning), a multi-image MLLM tailored for group-level aesthetics. We propose a progressive curriculum learning strategy and introduce a novel, model-agnostic *Visual Triager* to exploit the positional preference inherent in long-context multi-image reasoning.

4.1 CURRICULUM LEARNING FRAMEWORK

Training multi-image MLLMs for aesthetic reasoning is fundamentally more challenging than standard instruction tuning. Aesthetic visual understanding requires fine-grained perception of subtle details (e.g., exposure, focus, color grading), and existing open-source MLLMs, such as Qwen2.5-VL-Instruct (Bai et al., 2023b), suffer severe performance degradation over long visual sequences.

To overcome these dual bottlenecks, we strategically leverage our auxiliary and multi-group datasets to establish a progressive curriculum learning framework. Rather than immediately overwhelming the model with large image collections, the training schedule scales progressively in both task complexity and visual context length. The curriculum advances from simple foundational perception (Single-Photo Judgments and Pair-wise Comparisons) to *Single-Group Critiques*, and ultimately culminates in complex *Multi-Group Comparisons* and *Group-Reasoning MCQs*. This final stage requires the model to synthesize cross-image relationships, from low-level technicalities to high-

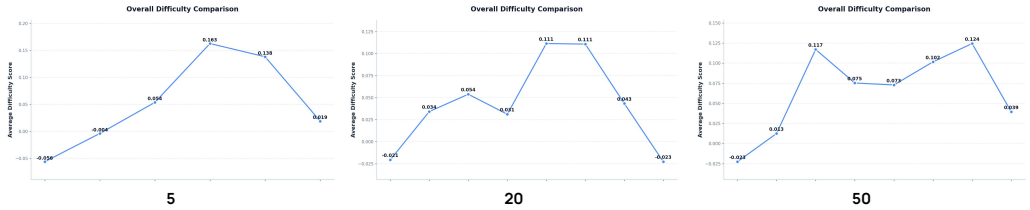


Figure 5: **The "Lost in the Middle" effect in multi-image aesthetic reasoning.** We plot the average perceptual difficulty score (y-axis) against the image's position in the input sequence (x-axis) for sequence lengths $L \in \{5, 20, 50\}$. Across all sequence lengths, the model exhibits pronounced positional sensitivity, struggling significantly more to evaluate images placed in the middle of the context compared to those positioned at the sequence boundaries.

level narratives, and contrast them across distinct groups. An overview of our training stages and corresponding data configurations is detailed in Appendix C (Tab. 5).

4.2 Visual Triager

Quantifying Positional Difficulty. In multi-image reasoning, an LMM may exhibit pronounced *positional sensitivity*: its perception of an image can vary depending on its position within the input sequence. To rigorously quantify this effect for aesthetic understanding, we define a position-dependent *perceptual difficulty* score based on the model's tendency to misjudge a visual attribute. Given an image $\boldsymbol{X}$ placed at position i in a sequence of length L, and a specific key phrase (attribute) t associated with an aesthetic topic T, we prompt the model with a binary question: whether $\boldsymbol{X}$ exhibits attribute t. Let ℓ_{yes} and ℓ_{no} denote the output logits corresponding to the *Yes* and *No* answers at the first decoding step. We compute the probability of the model rejecting the attribute via a binary softmax: $p_{\text{no}}(\boldsymbol{X}, t; i) = \frac{\exp(\ell_{\text{no}})}{\exp(\ell_{\text{yes}})+\exp(\ell_{\text{no}})}$. We define the perceptual difficulty of attribute t for image $\boldsymbol{X}$ at position i as $d_i(\boldsymbol{X}, t) \triangleq p_{\text{no}}(\boldsymbol{X}, t; i)$. A larger value indicates a higher tendency to reject the attribute, reflecting greater perceptual difficulty. For a topic T comprising a set of associated key phrases, we define the topic-level difficulty for image $\boldsymbol{X}$ at position i as the average difficulty across its phrases: $D_i(\boldsymbol{X}, T) = \frac{1}{|T|}\sum_{t\in T} d_i(\boldsymbol{X}, t)$. Let $\mathcal{X}$ denote the set of evaluated images. The expected difficulty of topic T at position i across the dataset is simply $D_i(T) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{X}\in\mathcal{X}} D_i(\boldsymbol{X}, T)$.

The "Lost in the Middle" Effect. We conduct this analysis on 100K images across all 17 aesthetic topics from VisGroup. We fix the input sequence length L and vary the target position $i \in \{1, \ldots, L\}$, computing $D_i(T)$. As illustrated in Fig. 5, across all tested sequence lengths ($L = 5$ to 50), the model consistently yields lower difficulty scores (i.e., performs more accurately) on images near the *beginning* and *end* of the sequence. Conversely, images in the *middle* suffer significant perceptual degradation. This phenomenon acts as a visual analogue to the "Lost in the Middle" effect observed in long-context text LLMs (Liu et al., 2024b).

Formulating the Visual Triager. This predictable positional bias motivates an elegant inference-time strategy: *dynamic image reordering*. By identifying difficult images and explicitly placing them at the boundary positions where the model is highly perceptive, we can construct a more robust reasoning pipeline. To implement this, we design a *Visual Triager* to predict the relative difficulty of assessing an image's aesthetics. Ideally, the overall difficulty of an image $\boldsymbol{X}$ under topic T for a sequence of length L is its expectation over all possible positions: $D(\boldsymbol{X}, T; L) = \mathbb{E}_{i\in\{1,\ldots,L\}}\Big[D_i(\boldsymbol{X}, T)\Big]$. Computing this explicitly is intractable at scale. However, we adopt the budget setting $i = 1, L = 1$ (evaluating a single image in isolation) as a highly reliable proxy. While varying i and L alters the absolute difficulty, empirical analysis (detailed in the Appendix) shows that the *relative* difficulty ranking $D_i(\boldsymbol{X}_1, T) - D_i(\boldsymbol{X}_2, T)$ between any image pair remains remarkably stable across different choices of (i, L). Thus, predicting the single-image difficulty preserves the relative difficulty ordering required for sorting.

Training and implementation details for the Visual Triager are provided in Appendix E.1.

Model	*Color & Tone*	*Lighting*	*Narrative*	*Emotion*	*Low-level Visuals*	*Composition*	*Post-processing*
InternVL3-78B	71.9	53.8	**89.5**	73.1	40.4	66.7	46.6
Qwen3.8-27B	71.9	65.4	73.7	65.4	57.7	75.0	62.1
Qwen2.5-72B	65.6	26.9	52.6	61.5	38.5	55.6	44.8
Qwen2.5-7B	62.5	30.8	52.6	57.7	42.3	52.8	32.8
Gemini 3.6	65.6	69.2	68.4	57.7	61.5	**80.6**	60.3
Grok 4.7	62.5	53.8	78.9	65.4	63.5	66.7	43.1
GPT-5.6 Sol	65.6	57.7	73.7	61.5	48.1	72.2	62.1
GPT-5.5	50.0	50.0	63.2	50.0	21.2	66.7	39.7
Opus 5	75.0	61.5	**89.5**	**76.9**	67.3	77.8	58.6
AVIS-7B (Ours)	**87.5**	**76.9**	73.7	**76.9**	**92.3**	75.0	**75.9**

Model	*Techniques*	*Culture*	*Technical Flaws*	*Aesthetic Style*	*Focus*	*Interaction*	**Overall**
InternVL3-78B	68.8	50.0	59.5	48.7	70.4	45.5	58.4
Qwen3.8-27B	**75.0**	37.5	62.0	59.0	70.4	54.5	64.5
Qwen2.5-72B	56.2	37.5	32.9	48.7	74.1	45.5	47.2
Qwen2.5-7B	62.5	50.0	36.7	43.6	55.6	45.5	45.1
Gemini 3.6	50.0	37.5	65.8	46.2	81.5	54.5	63.6
Grok 4.7	43.8	50.0	51.9	41.0	55.6	27.3	54.7
GPT-5.6 Sol	62.5	50.0	63.3	43.6	44.4	36.4	58.4
GPT-5.5	43.8	50.0	41.8	30.8	29.6	18.2	41.6
Opus 5	**75.0**	37.5	58.2	51.3	70.4	36.4	65.0
AVIS-7B (Ours)	62.5	**62.5**	**79.7**	**87.2**	**92.6**	**63.6**	**80.6**

Table 1: Results on VisGroup-Bench (Single). Best results in **bold**.

5 EXPERIMENTS

Experimental Setup. We build AVIS-7B by fine-tuning the Qwen2.5-VL-Instruct (7B) (Bai et al., 2025) backbone on our VisGroup dataset. Training is conducted on 8 A100 GPUs using DeepSpeed ZeRO-3 (details are provided in Appendix D).

Baselines & Benchmarks. Because existing specialized aesthetic MLLMs (Wu et al., 2024b;c) cannot process dozens of images simultaneously, we benchmark AVIS-7B against state-of-the-art general-purpose multi-image MLLMs. This includes proprietary models and leading open-source models. We evaluated the models on **VisGroup-Bench**.

Aesthetic Visual Understanding Tab. 1 summarizes performance on single-group understanding. AVIS-7B achieves state-of-the-art results with an overall accuracy of **80.6%**, outperforming the next best model (GPT-5.6 Sol at 58.4% and Opus 5 at 65.0%). Importantly, AVIS-7B demonstrates a notable improvement over its own base model, validating the efficacy of the VisGroup dataset and our curriculum learning strategy. The gains are most pronounced in specialized, low-level photographic categories that require fine-grained visual inspection. While proprietary models like Opus 5 remain competitive in high-level semantic categories requiring broad world knowledge (e.g., *Narrative, Emotion*), AVIS-7B demonstrates that a specialized 7B model can dominate generalist giants in professional aesthetic domains. In Tab. 2, we evaluate the substantially more challenging multi-group setting, which requires synthesizing patterns within groups and contrasting them across groups. AVIS-7B maintains its superiority, achieving an overall score of **79.5%**. This confirms that our framework successfully endows MLLMs with the deep, specialized cross-image reasoning capabilities required for professional aesthetic curation.

Effectiveness of VisGroup. In Appendix H, we present studies demonstrating the effectiveness of the different parts of VisGroup.

5.1 EFFECTIVENESS OF THE VISUAL TRIAGER

To evaluate the efficacy of our proposed test-time optimization, we assess AVIS on VisGroup-Bench with and without the Visual Triager. Once the Triager predicts the difficulty scores for individual images, we evaluate different reordering strategies across two distinct settings:

Reordering. *Single-Image Reordering*. For MCQs where all images belong to a single collection, we test three configurations for placing the most difficult images: (i) *Beginning*: Hardest images at the front of the sequence. (ii) *End*: Hardest images at the back of the sequence. (iii) *Alternating*: Distributing hard images between both boundaries (e.g., hardest at the beginning, second hardest at the end, third at the beginning, etc.). *Multi-Group Reordering*. For MCQs involving multiple collections (specifically two-group comparisons), altering the global group order would disrupt the logical flow of the prompt. Instead, we reorder images *within* their respective groups to maximize their proximity to globally advantageous positions. For the first group, we place difficult images

Model	*Color & Tone*	*Lighting*	*Narrative*	*Emotion*	*Low-level Visuals*	*Composition*	*Post-processing*
InternVL3-78B	68.4	53.1	58.3	74.5	51.8	46.1	40.6
Qwen3.8-27B	56.1	40.8	16.7	74.5	39.3	32.9	42.2
Qwen2.5-72B	54.4	42.9	41.7	74.5	28.6	34.2	35.9
Qwen2.5-7B	38.6	28.6	41.7	59.6	37.5	27.6	32.8
Gemini 3.6	38.6	61.2	33.3	61.7	44.6	50.0	42.2
Grok 4.7	64.9	51.0	66.7	63.8	57.1	51.3	53.1
GPT-5.6 Sol	71.9	57.1	79.2	66.0	51.8	57.9	60.9
GPT-5.5	57.9	63.3	87.5	51.1	44.6	53.9	56.2
Opus 5	73.7	71.4	79.2	70.2	58.9	67.1	**71.9**
AVIS-7B (Ours)	**77.2**	**87.8**	**95.8**	**95.7**	**92.9**	**75.0**	68.8

Model	*Techniques*	*Culture*	*Technical Flaws*	*Aesthetic Style*	*Focus*	*Interaction*	**Overall**
InternVL3-78B	56.7	54.5	38.8	50.0	43.5	54.2	51.8
Qwen3.8-27B	40.0	50.0	38.8	50.0	34.8	45.8	43.5
Qwen2.5-72B	60.0	63.6	35.3	44.7	34.8	58.3	44.2
Qwen2.5-7B	36.7	63.6	22.4	39.5	39.1	50.0	36.5
Gemini 3.6	56.7	22.7	30.6	44.7	43.5	45.8	44.5
Grok 4.7	63.3	54.5	36.5	39.5	60.9	58.3	53.4
GPT-5.6 Sol	53.3	59.1	44.7	50.0	34.8	54.2	56.8
GPT-5.5	60.0	54.5	45.9	42.1	52.2	37.5	53.3
Opus 5	70.0	63.6	51.8	57.9	65.2	66.7	65.7
AVIS-7B (Ours)	**80.0**	**95.5**	**57.6**	**86.8**	**87.0**	**75.0**	**79.5**

Table 2: Results on VisGroup-Bench (Multi). Best in **bold**, second best underlined.

Method	Single-Group		Multi-Group	Average
AVIS-7B	80.6		79.5	80.1
+ Visual Triager	Beginning	81.7 +1.1	**82.9** +3.4	**82.6** +2.5
	End	**82.2** +1.6		
	Alternating	81.9 +1.3		

Table 3: Effectiveness of Visual Triager on VisGroup-Bench. For single-group MCQs, we place more difficult images at the beginning, the end, or alternating between the two. For multi-group MCQs, we reorder images within each group to favor relatively advantageous positions.

toward the *front*; for the second group, we place them toward the *back*. This ensures hard images are placed as close to the absolute sequence boundaries as possible.

Results. As demonstrated in Tab. 3, applying the Visual Triager yields consistent and notable improvements across all configurations, boosting average performance by **+2.5%**. Interestingly, placing difficult images at the *End* slightly outperforms the *Beginning* and *Alternating* strategies for single groups. Crucially, the performance gains are significantly larger in the Multi-Group setting (**+3.4%**). This aligns with our hypothesis: multi-group comparisons intrinsically involve more images and longer contexts, making the model substantially more susceptible to positional bias. Consequently, the reordering becomes an increasingly vital optimization as the visual context scales. Computational efficiency measurements for the Visual Triager are provided in Appendix E.2.

The "Lost in the Middle" Phenomenon. The success of this simple triage-and-reorder step confirms the positional preferences in MLLMs. While this "Lost in the Middle" effect has been heavily studied in long-context text LLMs (Liu et al., 2024b), likely stemming from structural regularities in pretraining corpora where key information appears at document boundaries, it remains critically underexplored in multi-image contexts. Our findings suggest that MLLMs inherit these positional biases from their underlying LLM backbones. While the Visual Triager successfully mitigates this at test time, a comprehensive study to fundamentally eliminate positional preferences during MLLM pretraining remains an important direction for future research.

6 Conclusion

We introduce the first large-scale dataset for group-level aesthetic visual reasoning in photography, addressing the limitations of existing models and datasets focused on single-image assessment. By developing AVIS, a multi-image MLLM fine-tuned on the new VisGroup dataset, we demonstrate significant improvements in group-level aesthetic understanding. Our systematic analysis highlights the unique challenges of multi-image reasoning and provides valuable insights for advancing aesthetic evaluation in real-world photographic workflows.

AI USE STATEMENT

In this work, we used generative AI tools for data generation and to assist with parts of the coding. We have not used generative AI tools for other applicable tasks requiring disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for grammar checking. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We have read the Code of Ethics, adhere to it, and explicitly acknowledge this during the submission process.

REPRODUCIBILITY STATEMENT

We will release the model and dataset.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b. URL https://arxiv.org/pdf/2308.12966.

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.

Giovanna Castellano and Gennaro Vessio. Deep convolutional embedding for digitized painting clustering. In *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021. IEEE. doi: 10.1109/ICPR48806.2021.9412438.

Giovanna Castellano and Gennaro Vessio. A deep learning approach to clustering visual arts. *International Journal of Computer Vision*, 130:2590–2605, 2022. doi: 10.1007/s11263-022-01668-y.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. URL https://arxiv.org/abs/2411.18211v1.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023c.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, Yanlin Li, Lei Ren, Wei Chen, Zhiyuan Huang,

Mingjie Zhan, Xiaojie Wang, and Fangxiang Feng. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2506.04280. URL https://doi.org/10.48550/arXiv.2506.04280.

Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 1(3):269–288, 2005. doi: 10.1145/1083314.1083317.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Abhishek Dangeti, Pavan Gajula, Vikram Jamwal, and Vivek Srivastava. Enhancing artwork style clustering via neural representation re-alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 435–444. IEEE, 2025a.

Abhishek Dangeti, Pavan Gajula, Vivek Srivastava, and Vikram Jamwal. Style-based clustering of visual artworks and the play of neural style-representations. *arXiv preprint arXiv:2409.08245*, 2025b. URL https://arxiv.org/abs/2409.08245v2.

Yingying Deng, Fan Tang, Weiming Dong, Fuzhang Wu, Oliver Deussen, and Changsheng Xu. Selective clustering for representative paintings selection. *Multimedia Tools and Applications*, 78:19305–19323, 2019. doi: 10.1007/s11042-018-7139-4.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. AesExpert: Towards multi-modality foundation model for image aesthetics perception. *arXiv preprint arXiv:2404.09624*, 2024a.

Zhipeng Huang, Zhizheng Zhang, Yiting Lu, Zheng-Jun Zha, Zhibo Chen, and Baining Guo. Visualcritic: Making lmms perceive visual quality like humans. *arXiv preprint arXiv:2403.12806*, 2024b.

Takamu Kaneko and Keiji Yanai. Event photo mining from twitter using keyword bursts and image clustering. *Neurocomputing*, 172:143–158, 2016. doi: 10.1016/j.neucom.2015.02.081. URL https://doi.org/10.1016/j.neucom.2015.02.081.

Jiyeon Kim and Youngok Kang. Automatic classification of photos by tourist attractions using deep learning model and image feature vector clustering. *ISPRS International Journal of Geo-Information*, 11(4):245, 2022. doi: 10.3390/ijgi11040245. URL https://doi.org/10.3390/ijgi11040245.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025. URL https://arxiv.org/abs/2506.18883v1.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.07575*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint*, 2024a. Emails: liuhaowei2019@ia.ac.cn, cfyuan@nlpr.ia.ac.cn, {shuofeng.xhy, ym119608}@alibaba-inc.com.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models, 2024. URL https://arxiv.org/abs/2408.02718.

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

OpenAI. Gpt-4 technical report. In *arXiv*, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

J. C. Platt, M. Czerwinski, and B. A. Field. Phototoc: automatic clustering for browsing personal photographs. In *Proceedings of the IEEE Conference*. IEEE, December 2003. Presented at the IEEE Conference on Information Visualization, 15–18 Dec 2003.

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

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models, 2025. URL https://arxiv.org/abs/2503.13792.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2025. URL https://arxiv.org/abs/2410.03290v2.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-BENCH: A benchmark for general-purpose foundation models on low-level vision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024a. URL https://q-future.github.io/Q-Bench.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024c.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Siyu Xia, Hong Pan, and A. K. Qin. Face clustering in photo album. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, 2014. IEEE. doi: 10.1109/ICPR.2014.490.

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint*, 2024.

Zhipeng Zhong, Fei Zhou, and Guoping Qiu. Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

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

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

A Extended Related Works

As discussed in the main text, constructing datasets for group-level aesthetic visual understanding requires identifying groups of images with detailed aesthetic annotations. However, existing datasets for aesthetic visual assessment either provide extensive photo critiques without any group-level structure (Zhong et al., 2023; Qi et al., 2025; Wu et al., 2024b), or offer grouping information but lack the fine-grained aesthetic labels necessary for group-level analysis, such as Unsplash[3]. There also exist several lines of work related to forming image groups, though they are not directly aligned with our goal.

One line focuses on style-based clustering of artworks (Castellano & Vessio, 2021; Dangeti et al., 2025b; Castellano & Vessio, 2022; Deng et al., 2019; Dangeti et al., 2025a). These clusters are typically constructed using style as the primary criterion, often derived automatically from clustering algorithms. In contrast, our image groups consist of photographic works and are organized according to a highly diverse set of aesthetic criteria, including authorship, subject matter, techniques, expressive intent, affect, narrative structure, and other aesthetic topics. Each topic further expands into hundreds or even thousands of keywords, and aesthetic style constitutes only one of many grouping dimensions. Such rich and domain-specific grouping cannot be reliably obtained from visual features alone, which motivates our carefully designed LLM-based grouping pipeline. Moreover, each image in our dataset is paired with detailed, human-annotated photo critiques, which existing datasets do not provide and are essential for producing group-level critiques.

Another line of work performs photo clustering (Xia et al., 2014; Platt et al., 2003; Cooper et al., 2005; Kim & Kang, 2022; Kaneko & Yanai, 2016), generally focusing on event-based clustering or organizing personal photo albums. These approaches similarly cannot support the fine-grained, aesthetic-topic-level grouping we require, nor do they include detailed critique for individual photos, which are necessary for group-level critique construction.

B Auxiliary Data Construction

In addition to the main group-level reasoning data, we generate several forms of auxiliary data, including single-photo judgments, pairwise comparisons, and multi-photo referring tasks, which are jointly trained with the group-level data.

B.1 Single-Photo Judgment

It is a binary QA task where the model determines whether a given photo satisfies a specific key phrase under an aesthetic topic (answering "yes" or "no"). For each image, there are 17 aesthetic topics, and in each aesthetic topic, there can be multiple key phrases. Therefore, for one image, there can be tens of `(image, aesthetic topic, key phrase)` triplets. When fully expanded, the DPC photo critique dataset (obtained from DPC 2022) contains roughly ten million such triplets, from which we randomly sampled about 1.5% of the data to form the single-photo judgment dataset. To construct negative samples, we randomly selected half of the data, replaced each key phrase with a new key phrase that is semantically opposite or mutually exclusive to the original one, and changed the answer to `No`. An example is shown in Fig. 6. Note that, to obtain mutually exclusive phrases for each key phrase, we instructed `LLaMA-3.3-70B-Instruct` to generate contrasting alternatives for the source phrases.

B.2 Pair-wise Comparison

It is a two-photo QA task constructed from pairs of photos sampled under the same aesthetic topic but representing mutually exclusive key phrases (e.g., under *Lighting Characteristics*: "Natural Lighting" vs. "Rembrandt Lighting"). The model is asked which photo better fits a given key phrase, and it responds with the index of the correct photo. To construct this part of the data, we first randomly sampled `(image, aesthetic topic, key phrase)` triplets from the set described above, and for each triplet, we find another triplet with the same aesthetic topic but contrasting key phrase to form pairs. This step is critical to avoid selecting two key phrases that convey

[3] `https://unsplash.com/data`



Question: Does the image reflect serenity in terms of Emotion?

Answer: Yes

Figure 6: An example of single-photo judgment.

similar meanings, which could confuse the model. In practice, given a source key phrase, we employ `LLaMA-3.3-70B-Instruct` to assess each candidate key phrase and identify whether it shares the same semantics as the source phrase. We only retain key phrases that are semantically opposite or mutually exclusive to the source for pairing. We then randomly selected one triplet in the qualified pair as the ground truth and recorded its key phrase. Finally, we asked which of the two images better reflects the given (recorded) key phrase. The size of this subset is also approximately 1% of the total number of triplets, roughly 100K samples. An example is shown in Fig. 7.

Negative Key Phrase Mining. As mentioned above, given a source key phrase, we employ the LLM to evaluate each candidate key phrase and determine whether it shares the same semantics as the source. In practice, however, this approach is extremely inefficient, as every key phrase must be compared against all remaining phrases using the LLM, resulting in a time complexity of $\mathcal{O}(n^2)$. Since an aesthetic topic typically contains thousands of key phrases, this would generate tens of millions of LLM comparison pairs and become prohibitively time-consuming.

To address this, we carefully design a more efficient strategy. For each source key phrase, we first instruct the LLM to generate ten mutually exclusive or contrasting key phrases. We then compute their embeddings using a sentence transformer and take the average to form a *negative embedding* for the source key phrase. This negative embedding is used to rank all key phrases within the same aesthetic topic based on similarity, with phrases closer to the negative embedding ranked higher. We then select the top N candidates and let the LLM make the final determination of whether each candidate qualifies as a negative phrase for the source phrase. In practice, we set $N = 400$, which reliably covers all negative phrases in almost all cases. This reduces the overall complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, yielding a substantial efficiency improvement. Note that we use `LLaMA-3.3-70B-Instruct` in the negative key phrase mining process.

B.3 Multi-Photo Referring

We also extend the pair-wise comparison idea to a multi-photo context. For each sample, we randomly select a key phrase from an aesthetic topic and gather 1–15 photos that satisfy it as positive samples. We then randomly select several mutually exclusive key phrases and corresponding photos (1–15) as negatives, mix them together, and require the model to identify the subset of positive photos that match the target key phrase. The first two datasets aim to improve the model's basic understanding of fundamental aesthetic features through simpler tasks, while the multi-photo referring

task enhances its basic multi-image understanding ability via referring in aesthetic contexts. As it is more challenging, we sampled approximately 430K data for this task.



Question: Select the image that more strongly reflects soft HDR in terms of its Post-processing Style

Answer: 1

Figure 7: An example of pair-wise comparison.

C Dataset Construction

Category	Author	Challenge	Gallery	Galleries
Count	20613	3034	69	1952
Average Photos	22.05	149.78	7859.94	232.80

Table 4: Dataset Statistics of DP Challenge (from DPC2022 (Zhong et al., 2023)).

C.1 VisGroup

Single-Group Critique. Starting from the groups obtained in Section 3.1, we first performed a preliminary filtering step: groups containing fewer than 10 images were discarded. For groups with more than 10 images, we applied the sampling procedure described in Section 3.2 of the main text to obtain multiple image subsets without replacement. We then processed these image groups using `LLaMA-3.3-70B-Instruct`, following the workflow illustrated in Fig. 4 of the main

Stage	Datasets	#Data
Stage 1	Single-Photo Critique	280K
Stage 2	Single-image Judgment, Pair-wise Comparison, Single-Group Critique, Single-group Critique MCQ	1.1M
Stage 3	Image Referring, Multi-Group Critique, Multi-Group Critique MCQ, Multi-Group Reasoning MCQ	96K

Table 5: Training curriculum across stages. Each stage integrates progressively more complex datasets to improve multi-image aesthetic reasoning.

Dataset	Size	Avg. #Img	Avg. Ans. Length
Single-group			
Single-group Critique	840K	24.10	29.63
Single-group Critique MCQ	100K	25.44	-
Multi-group			
Multi-group Critique	268K	59.54	51.90
Multi-group Critique MCQ	122K	42.61	-
Multi-group Reasoning MCQ	148K	50.00	-
Auxiliary Data			
Single-image Judgment	140K	1.00	-
Pair-wise Comparison	100K	2.00	-
Image Referring	430K	23.33	9.243

Table 6: Overview of the VisGroup dataset. *Avg. #Img* is the average number of images per training sample, and *Avg. Ans. Length* is the average length of answers in the training set.

text. Subsequently, we applied `LLaMA-3.3-70B-Instruct` to filter these samples, asking the LLM to identify whether a generated group-level critique contained any internal inconsistencies, for example, claiming that the same image group primarily reflects two mutually contradictory features. While a group may indeed exhibit multiple characteristics for an aesthetic topic, it typically has only one dominant feature (if there is any common feature for this topic), and critiques violating this principle were removed. We ultimately obtained 840K single-group critiques.

Multi-Group Critique. This processing pipeline is described in detail in Section 3.2 and Fig. 4 of the main text. The LLM we used is `LLaMA-3.3-70B-Instruct`. Following a procedure similar to the filtering of single-group critiques, we employed `LLaMA-3.3-70B-Instruct` to remove critiques containing logically self-contradictory statements. The final size of this dataset is 268K.

Aesthetic Multiple Choice Questions. Finally, we generated Multiple-Choice Questions (MCQs) based on the group-level critique data. For the Group-critique MCQs, we provided either single-group or multi-group critiques to `LLaMA-3.3-70B-Instruct` and prompted it to generate a high-quality MCQ for each critique. We then performed a filtering step by feeding both the generated MCQ and the original critique back into the LLM, asking it to judge whether the MCQ was faithful to the critique, namely, whether the ground-truth answer was guaranteed to be correct and whether the distractors were guaranteed to be incorrect. MCQs that did not meet the criteria were removed. In the end, we obtained 100K Single-group Critique MCQs and 122K Multi-group Critique MCQs.

We further created 148K Multi-group Reasoning MCQs, where a source image group, originating from a particular photographer (sampled from author groups) or characterized by a specific aesthetic feature (sampled from aesthetic key-phrase groups), is presented along with four candidate groups. One of the candidate groups (the ground-truth answer) comes from the same photographer or shares the same characteristic as the source group, while the others originate from different photographers or characteristics. This task is designed to evaluate whether a model can understand the distinctive attributes of image groups and perform reasoning and comparison across different groups.

C.2 VisGroup-Bench

To evaluate the group-level understanding ability of MLLMs, we further construct a comprehensive benchmark, VisGroup-Bench, also derived from the DP Challenge. Following the same procedure described above, we form image groups based on individual photo critiques and meta information, from which we obtain group-level descriptions. Subsequently, we generate aesthetic multiple-choice questions with the process demonstrated in the previous section. Finally, we evaluate our method on 1,291 single-group and multi-group multiple-choice questions, comprising 429 single-group questions and 862 multi-group questions (595 main-evaluation questions and 267 Photographer Identity questions).

Remark. Model generalizability is crucial. VisGroup-Bench is carefully designed to assess whether MLLMs finetuned on VisGroup can really improve their group-level aesthetic visual understanding capabilities, thereby demonstrating their capacity to generalize. DP Challenge itself is highly diverse, encompassing thousands of photographic themes that cover nearly all major genres of pho-

Hyperparameter	Value
batch size (per device)	2
eval batch size	4
gradient accumulation steps	8
num train epochs	1
learning rate	2e-7
weight decay	0
warmup ratio	0.03
max grad norm	1
lr scheduler	cosine
model max length	32768
gradient checkpointing	True
precision	bf16
dataloader num workers	4
max pixels	50176
min pixels	784
DeepSpeed stage	3

Table 7: Training hyperparameters used in finetuning.

tography. Moreover, the images are contributed by hundreds of thousands of registered users, greatly enriching the diversity and mitigating overfitting to specific data distributions. Beyond this inherent diversity, we take an additional step in constructing the benchmark by explicitly reserving a subset of authors and aesthetic features that **never** appear in the training set as evaluation data. This allows us to examine not only whether the model can effectively learn common aesthetic concepts from VisGroup and generalize them to unseen images (which can be achieved through a standard train–test split based on images, where however, authors and aesthetic key phrases may overlap), but also examine whether it can extend its learned aesthetic knowledge to unseen aesthetic concepts and authors.

D Experiment Details

D.1 Experimental Configuration

We build AVIS-7B upon Qwen2.5-VL-7B-Instruct (Bai et al., 2025) on 8 A100 GPUs with ZeRO3. In Tab. 7, we report the detailed experimental settings for the supervised finetuning.

D.2 Referring to Images

In multi-image understanding and video understanding with MLLMs, two common strategies are used to refer to specific images or frames in the input: inserting special temporal tokens between images (Chen et al., 2024; Li et al., 2025; Wang et al., 2025), or directly attaching numerical identifiers (Wu et al., 2025) to the images themselves. Recent works have shown the effectiveness of the latter approach across various MLLMs, and we adopt this simple yet effective strategy to annotate input images. Specifically, we place a red numerical label in the upper-left corner of each image, marking the first image as 1, the second as 2, and so on, thereby better indicating the positional index of each image to the MLLM.

E Visual Triager

E.1 Training and Implementation Details

We construct the training data for the Visual Triager using 28K images from the Single-Photo Critique subset. We first assign continuous difficulty labels by computing p_{no} for each (image, aesthetic topic, key phrase) triplet, yielding 673K annotated triplets. We aggregate these to form an overall image-level difficulty score $D(\boldsymbol{X}) \in [0, 1]$ for each image. We fine-tune the lightweight Qwen2.5-VL-3B-Instruct as our Visual Triager. We treat the continuous difficulty scores as text

for standard supervised fine-tuning using the prompt: *"Assess the difficulty of aesthetic evaluation for this image on a 0–1 scale."* The model is trained via standard next-token prediction, formatted as: **User:** <IMG> + prompt; **Assistant:** {score}. At test time, the Visual Triager independently processes each image in the input set, outputting a scalar difficulty score. The images are subsequently reordered based on these scores and passed to the LLM, significantly mitigating positional bias without requiring complex architectural modifications.

E.2 Computational Efficiency

A critical requirement for any test-time optimization is that it does not introduce prohibitive latency. Because our Triager is built on the lightweight Qwen2.5-VL-3B architecture, it natively supports highly efficient batched inference (processing one image per batch independently). For a rigorous comparison, we measure inference time on a sequence of 50 images where both the Triager and AVIS output exactly 1 token (e.g., an MCQ answer). Under these conditions, the Triager requires only $\mathbf{0.41}\times$ the inference time of AVIS. More importantly, in real-world scenarios where AVIS generates detailed group descriptions of length T tokens, the Triager's relative overhead shrinks dramatically. Because the Triager only predicts a single scalar once per image, its relative latency decreases linearly with T, falling below $\mathbf{0.05}\times$ the time of AVIS for generated texts longer than 10 tokens. This shows that the Visual Triager is a scalable, negligible-cost addition to the pipeline.

E.3 Choice of i and L

We set $i = L = 1$ as the default configuration for estimating image difficulty. This choice is motivated by theoretical analysis and empirical observations.

Varying i and L mainly affects the *absolute* difficulty value but has limited impact on the *relative* difficulty ordering among samples. For two images $(\mathbf{X}_1, \mathbf{X}_2)$, the difference of their overall difficulty under configuration (i, L) is

$$D(\mathbf{X}_1, T; L) - D(\mathbf{X}_2, T; L) = \frac{1}{L} \sum_{i=1}^{L} \big(D_i(\mathbf{X}_1, T) - D_i(\mathbf{X}_2, T)\big) \tag{1}$$

If the pairwise difficulty differences $D_i(\mathbf{X}_1, T) - D_i(\mathbf{X}_2, T)$ remain relatively stable across different (i, L) choices, the resulting ranking of relative difficulty will also be preserved. Empirically, using the relative difficulty ranking obtained with $L = i = 1$ as a reference, the rankings under other choices of (i, L) exhibit little variation, as indicated by SRCC values close to 1. As shown in Tab. 8, the relative ordering of image difficulty remains highly consistent across different parameter settings. Since larger (i, L) values introduce additional computation but bring minimal change to the ranking, we adopt $i = L = 1$ as a budget-friendly choice.

(i=L=1): 1 / +2.3	(i=1, L=10): 0.93 / +2.4	(i=5, L=10): 0.91 / +2.2	(i=L=10): 0.92 / +2.3

Table 8: **S**pearman **R**ank **C**orrelation **C**oefficient (SRCC) shows relative image difficulty order difference. For (i, L), we show SRCC (i=L=1 as reference) scores and the average gains on benchmark with Triager in Blue. i=L=1 is a budget choice.

F Image Difficulty Distribution

We used AVIS-7B to assign a difficulty label to each `(image, aesthetic topic, key phrase)` triplet in the training data for the visual triager. In total, 673K triplets were annotated. These labels were then aggregated at the image level for each image, yielding difficulty estimates for 28K images. The resulting distribution is shown in Fig. 8.

G Discussion on Single-image Aesthetic Understanding Model

Adapting Single-Image Aesthetic MLLM to Multi-Image Reasoning. A simple way to adapt such MLLM to multi-image setting is subset inference with majority voting. We adopt Co-Instruct (Wu et al., 2024c) for this study. For each question, we split the image group into subsets

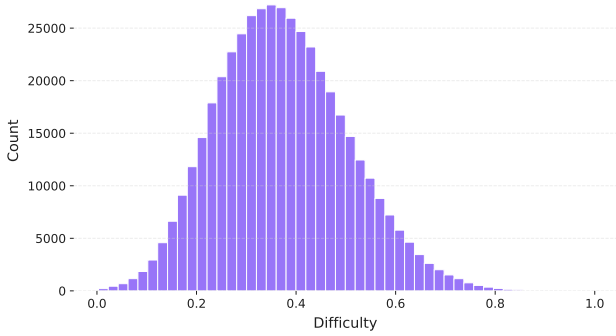


Figure 8: Image difficulty distribution.

Method	Metric
Multi-Group Evaluation	
Full Dataset	79.5
w/o Auxiliary Data	77.4
w/o Multi-Group Data (stage 3)	71.2
Single-Group Evaluation	
Full Dataset	80.6
w/o Auxiliary Data	78.4
w/o Group Data (stage 2, 3)	63.5
w/o Single Image Data (stage 1)	57.1

Table 9: Performance comparison across different training subsets.

of up to four images (the maximum input size of Co-Instruct (Wu et al., 2024c)), run inference on each subset independently, and aggregate the predictions via majority voting. It achieves an average accuracy of 32.7% over the single and multi group tests, notably inferior to the baselines and ours.

H Effectiveness of VisGroup

As shown in Tab. 5, our training data are organized into stages. Our goals from Stage 1 to Stage 3 progress from single-image aesthetic visual understanding to single-group aesthetic visual understanding and finally to multi-group aesthetic visual understanding. To strengthen the model's multi-image understanding capabilities, we additionally incorporate auxiliary data during training as previously discussed.

As shown in Tab. 9, each component of the training data contributes noticeably to the model performance. The group-level data in Stages 2 and 3 substantially enhance the model's ability in single-group understanding, while the Multi-Group Data in Stage 3 notably strengthen the model's capability in multi-group understanding. We also observe that, although Stage 1 does not involve any multi-image tasks, the model's single-group evaluation score still increases from 57.1 to 63.5 after this stage. This gain can be attributed to the aesthetic knowledge learned from single-image critiques in Stage 1.

I Photographer Identity

Model	Accuracy (%)
InternVL3-78B	28.5
Qwen3.8-27B	27.7
Qwen2.5-72B	29.2
Qwen2.5-7B	26.6
Gemini 3.6	53.9
Grok 4.7	60.7
GPT-5.6 Sol	<u>68.2</u>
GPT-5.5	61.0
Opus 5	67.4
AVIS-7B (Ours)	**78.7**

Table 10: Results on the Photographer Identity subset of VisGroup-Bench (267 multi-group queries). Best in **bold**, second best <u>underlined</u>.

We separately evaluate photographer identity on 267 multi-group queries from VisGroup-Bench. For each query, the model is given a source image group and four candidate groups and must select the candidate group captured by the same photographer as the source, following the Group-Reasoning MCQ formulation in Section 3.2. Tab. 10 reports the accuracy of each model.