# Fine-grained Image Aesthetic Assessment: Learning Discriminative Scores from Relative Ranks

Supplementary Material

## 8. FGAesthetics

In this section, we provide supplementary details on the construction of FGAesthetics, including: (1) detailed source distribution statistics, (2) implementation procedures of series refinement and rank calibration, and (3) prompt engineering methodology for comparative text generation.

### 8.1. Source Distribution of Data Collection

Tab. 6 presents the detailed source distribution of FGAesthetics. We collect 106,632 images from eight diverse sources spanning three distinct categories: natural images [5, 49] , AI-generated content (AIGC) [9, 14, 20, 52], and cropping-based datasets [40, 50]. Through rigorous series refinement and rank calibration, the final dataset contains 32,217 images organized into 10,028 series, yielding an average of 4.47 pairs per series for fine-grained aesthetic discrimination. This diverse composition ensures that FGAesthetics encompasses comprehensive fine-grained aesthetic dimensions: natural photography quality, AI generation aesthetics, and compositional variations.

Table 6. Source Distribution and Statistics of FGAesthetics

| Sources | Collected Img Count | After Series Refinement | After Rank Calibration | Series Count | Pair Count | Avg. Pairs per Series |
|---|---|---|---|---|---|---|
| SPS [5] | 15,543 | 12,522 | 12,522 | 4,755 | 12,558 | 2.64 |
| LSVQ [49] | 2630 | 565 | 556 | 150 | 636 | 4.24 |
| Pick-a-pic [20] | 13,917 | 3,751 | 3,719 | 1,685 | 2,115 | 1.26 |
| Q-Eval-100K [52] | 11,430 | 418 | 372 | 180 | 196 | 1.09 |
| Midjourney [14] | 11,352 | 2,815 | 2,599 | 909 | 2,124 | 2.34 |
| NIGHTS [9] | 9,822 | 1,944 | 1,895 | 647 | 1,590 | 2.46 |
| CPC [40] | 29,348 | 8,485 | 8,480 | 1,310 | 21,242 | 16.22 |
| GAIC [50] | 12,590 | 2,088 | 2,074 | 392 | 4,402 | 11.23 |
| **Total** | **106,632** | **32,588** | **32,217** | **10,028** | **44,863** | **4.47** |

### 8.2. Details of Series Refinement

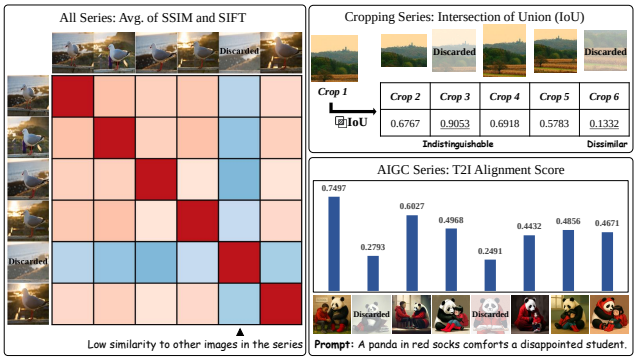
**Metrics Filtering.** Fig. 6 illustrates the detailed procedure of Metrics Filtering. We employ both generic and domain-specific measures to screen the data. First, generic measures compute the average SSIM and SIFT scores between each image and all others within its series. We then rank all images across the entire dataset and filter out the bottom 30% to exclude obvious outliers. For cropping series, we further calculate Intersection over Union (IoU) between crop frames and remove images with IoU $> 0.8$ (indistinguishable) or IoU $< 0.2$ (dissimilar). For AIGC series, we leverage original human-annotated T2I alignment scores to filter images severely misaligned with text prompts.

**MLLMs Checking using Gemini.** Given the robust contextual understanding capabilities of MLLMs, we also employ Gemini-2.5-pro [37] to validate the filtered series. This is achieved by presenting all images within a series to the model, guided by a carefully crafted prompt that requires a structured output for systematic parsing:

*# system: You are a rigorous image series filtering expert. Your task is to filter a given series of images and identify a subset that adheres to the specified criteria.*

*# user: Your evaluation process is governed by the following criteria. Intra-group Similarity: Within the provided image series, identify the largest subset of images that are similar in theme, style, composition, and subject matter. They should appear as close variations of the same concept. If an image's theme significantly deviates from the other images, it must not be included in the final similar group. Your output must be a formatted LIST: "selected_uids": ["<uid_a>", "<uid_b>", ...].*

**Human Qualification.** Following the automated filtering stages, comprehensive human qualification is implemented to ensure final data quality. All series undergo manual review by five human annotators via the annotation platform shown in Fig. 7. Samples that fail to meet fine-grained aesthetic criteria are discarded, ensuring that only high-quality series proceed to the aesthetic ranking annotation stage.




(a) Generic measures (b) Domain-specific measures

Figure 6. Detailed Procedure of Metrics Filtering. (a) Generic measures (average of SSIM and SIFT) exclude outliers within visually similar photo series. (b) Domain-specific measures (IoU for cropping series, T2I alignment scores for AIGC series) effectively identify nearly identical and dissimilar images.

### 8.3. Details of Rank Calibration

Based on the refined data, we perform pairwise comparisons within each series to obtain global aesthetic rankings. Specifically, for a series of length $n$, a total of $C_n^2$ image pairs are generated. Each pair is evaluated by 10 trained an-

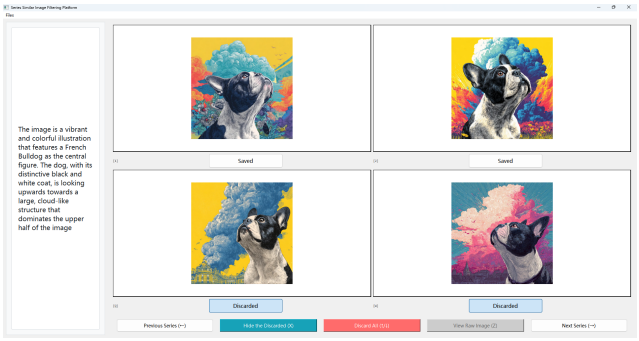



Figure 7. Annotation platform for Human Qualification. The platform displays all images from a single series on one page, enabling annotators to identify and filter outliers.

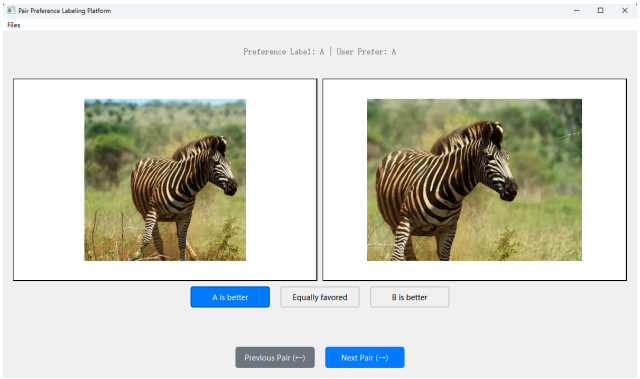



Figure 8. Annotation platform for Pairwise Comparison. The platform displays two images side-by-side for aesthetic comparison. Annotators make a forced-choice selection from three options: 'A is better', 'B is better', or 'Equally favored'.

notators for aesthetic comparison, determining which image is superior or indistinguishable, with the annotation platform shown in Fig. 8.

A total of 15 human evaluators (12 male, 3 female), all holding graduate degrees in computer science and actively engaged in IAA research, participated in the comprehensive annotation process for FGAesthetics. To mitigate potential biases from content variation, the annotation workload was systematically divided into three distinct sessions, each dedicated to one of the primary source categories: Natural, AIGC, and Cropping. The entire human annotation phase, which includes both the series refinement described before and the pairwise aesthetic comparisons, spanned 50 days.

### 8.4. Prompt Engineering for Reasoning Generation

Comparative textual descriptions are utilized to guide the visual model toward fine-grained aesthetic differences, enabling the CTAlign component of FGAesQ. To obtain these descriptions while minimizing human annotation burden, we leverage the robust reasoning capabilities of MLLMs.


Figure 9. Word clouds visualizing most frequent rationales for MLLM-generated comparative textual descriptions.

The detailed prompt template is provided below. Note that for AIGC and Cropping series, *Original Images* and *Shared Text Prompts* serve as additional contextual information.

*# system: You are an expert in digital art and aesthetics. You will be provided with a [Original Image, Shared Text Prompt] for context, two images, [Image A] and [Image B], and an aesthetic preference probability for Image A, [prob_A_preferred].*

*# user: Your evaluation should be based on the concepts in the [Original Image, Shared Text Prompt], assessing creative execution, aesthetics, and other aspects of various aesthetic attributes. Your task is to generate a single string where two sentences are separated by "||". The first sentence describes Image A, and the second describes Image B, creating an implicit but powerful contrast.*

*You must follow these crucial rules:*

1. *The sentences must not contain "Image A," "Image B," or similar direct labels; the comparison must be achieved through contrasting descriptions.*
2. *Each sentence must be concise (under 30 words) and use strong, comparative language (e.g., "far more refined," "lacks the depth," "pales in comparison").*
3. *Wording must not refer to the creation process (e.g., "prompt," "crop"); focus only on comparing the final images.*

*Your output must be a formatted JSON object and must not contain any explanatory text outside of the JSON format. The JSON structure is as follows:*

```
{
    "description": <Your description>,
    "prob_A_preferred": <the preference probability for Image A>
}
```

*For example, if the preference probability for Image A is 0.167 (B is preferred), your output should be:*

```
{
    "description": "This composition feels somewhat loose and the subject more distant, weakening the overall emotional
```

Table 7. Backbone ablation results on fine-grained and coarse-grained IAA tasks. For fine-grained evaluation on FGAesthetics, we report pair-level local discrimination using *Acc* and *F1*, and series-level ranking quality using *s-Acc* and *s-SRCC* across three source categories (Natural, AIGC, Cropping). *Pair* and *Series* represent category-averaged values of (*Acc*+*F1*)/2 and (*s-Acc*+*s-SRCC*)/2, respectively. For coarse-grained evaluation, we report SRCC and PLCC on AVA [26].

| Backbone | Natural | | | | AIGC | | | | Cropping | | | | Fine-grained | | Coarse-grained | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Pair* | *Series* | SRCC | PLCC |
| ViT-B/16 | **0.779** | **0.779** | **0.753** | **0.729** | **0.709** | **0.707** | **0.561** | **0.482** | **0.774** | **0.773** | 0.488 | **0.590** | **0.753** | **0.600** | **0.770** | **0.781** |
| ViT-B/32 | 0.711 | 0.711 | 0.634 | 0.503 | 0.653 | 0.651 | 0.445 | 0.301 | 0.740 | 0.740 | **0.491** | 0.548 | 0.701 | 0.487 | 0.747 | 0.760 |

Table 8. Out-of-Distribution generalization performance. OOD generalization is evaluated by training FGAesQ with one source category excluded and testing on fine-grained (all three categories) and coarse-grained (AVA [26]) benchmarks.

| Backbone | Natural | | | | AIGC | | | | Cropping | | | | Fine-grained | | Coarse-grained | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Acc* | *F1* | *s-Acc* | *s-SRCC* | *Pair* | *Series* | SRCC | PLCC |
| FGAesQ | 0.779 | 0.779 | 0.753 | 0.729 | 0.709 | 0.707 | 0.561 | 0.482 | 0.774 | 0.773 | 0.488 | 0.590 | 0.753 | 0.600 | 0.770 | 0.781 |
| w/o Natural | **0.745** | **0.745** | **0.710** | **0.630** | 0.704 | 0.703 | 0.569 | 0.455 | 0.771 | 0.771 | 0.488 | 0.583 | 0.740 | 0.573 | 0.770 | 0.782 |
| w/o AIGC | 0.767 | 0.766 | 0.720 | 0.675 | **0.687** | **0.685** | **0.542** | **0.425** | 0.765 | 0.765 | 0.482 | 0.577 | 0.739 | 0.570 | 0.776 | 0.786 |
| w/o Cropping | 0.765 | 0.764 | 0.727 | 0.678 | 0.699 | 0.698 | 0.552 | 0.438 | **0.755** | **0.755** | **0.462** | **0.558** | 0.739 | 0.569 | 0.772 | 0.783 |

```
impact.||By contrast, this one uses a much
tighter and more intimate focus, creating a
significantly more compelling narrative.",
    "prob_A_preferred":  0.167
  }
```

*Notably, the Comparison Direction of these two descriptions must be consistent with prob_A_preferred, that is, there cannot be a situation like prob_A_preferred is less than 0.5 but description A is more positive than description B.*

To validate the quality of generated descriptions, we conduct a human evaluation. Three independent annotators assess 100 randomly sampled descriptions for plausibility, achieving an agreement rate of 93%. This confirms that our prompt engineering approach effectively generates meaningful and accurate explanations capturing fine-grained aesthetic differences. In Fig. 9, we further visualize the word cloud of MLLM-generated descriptions, where explicit comparative terms (e.g., "more", "superior", "less") dominate the vocabulary.

# 9. FGAesQ

Following the implementation details and extensive experiments described in the paper, we present supplementary ablation studies to further validate the design choices of FGAesQ. Specifically, we investigate the impact of backbone architecture, Out-of-Distribution (OOD) generalization capability, and different DiffToken configurations.

## 9.1. Ablation of Backbone

We first investigate the impact of backbone architecture by replacing ViT-B/16 with ViT-B/32, a variant with a larger patch size. As shown in Tab. 7, ViT-B/16-based FGAesQ consistently outperforms ViT-B/32 across most fine-grained and coarse-grained evaluations. The performance gap can be attributed to the smaller patch size (16×16 vs. 32×32) of ViT-B/16, which enables the model to capture more granular visual details. This is crucial for discerning subtle aesthetic differences required for both fine-grained ranking and coarse-grained scoring.

## 9.2. Out-of-Distribution (OOD) Generalization

We further conduct an Out-of-Distribution (OOD) generalization analysis of FGAesQ, as illustrated in Tab. 8. Specifically, we evaluate FGAesQ trained with one source category excluded and test its performance on both fine-grained and coarse-grained IAA tasks. The results reveal that performance degradation is most pronounced on the test set corresponding to the excluded training category. For instance, when trained without Natural data, the model shows the largest performance drop on Natural evaluation (from 0.779/0.753 to 0.745/0.710 for Acc/s-Acc), while maintaining relatively stable performance on AIGC and Cropping. This pattern holds consistently across all three categories, demonstrating the distinct nature of each data source. Coarse-grained performance remains largely stable across all training configurations, with SRCC and PLCC consistently above 0.77 and 0.78, respectively. Notably, excluding AIGC data even leads to a slight improvement in coarse-grained metrics, which can be attributed to the fact that AVA predominantly consists of natural images.

## 9.3. Ablation of DiffToken

As introduced in the main paper, the Difference-preserved Tokenization (DiffToken) module is governed by two key hyperparameters: the difference localization patch size and the percentile parameter $p$, which determines the proportion of aesthetics-decisive regions. We conduct an ablation

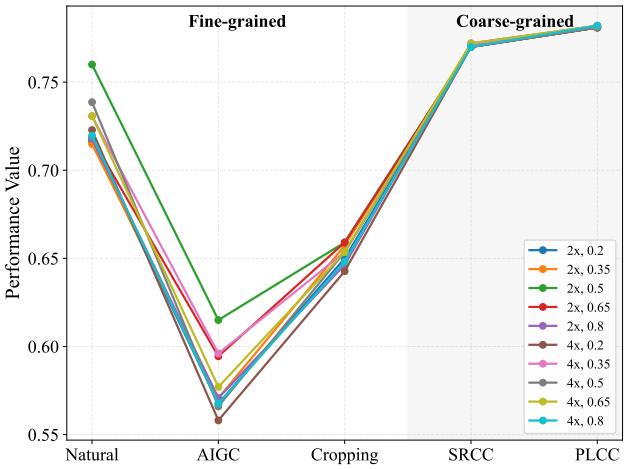



Figure 10. Ablation study on DiffToken configuration. Performance is evaluated across different difference localization patch sizes and percentile thresholds for identifying aesthetics-decisive regions. Fine-grained metrics show pair- and series-averaged performance across Natural, AIGC, and Cropping categories. Coarse-grained metrics report SRCC and PLCC on AVA [26].

study evaluating two patch sizes, 32×32 (2× ViT patch size) and 64×64 (4×), combined with five percentile values: $p \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, as illustrated in Fig. 10.

The results yield several key insights. First, the optimal configuration is 32×32 with $p = 0.5$, achieving the best overall performance across all tasks. Second, performance exhibits a clear inverted-U pattern with respect to $p$, peaking around 0.5 and degrading at extremes. A small $p$ (e.g., 0.2) fails to capture sufficient aesthetically-decisive regions, while too large $p$ (e.g., 0.8) forces aggressive downsampling of non-decisive regions to meet the token constraints, losing critical global compositional information. Third, the finer 32×32 localization consistently outperforms 64×64, as smaller patches more effectively detect subtle local variations crucial for fine-grained IAA. Finally, coarse-grained performance (SRCC/PLCC on AVA) remains stable across all configurations, confirming that DiffToken's hyperparameters primarily affect fine-grained discrimination while having negligible impact on absolute aesthetic assessment.

## 10. Additional Visual Examples

In Fig. 11, we present additional visualization examples demonstrating FGAesQ's fine-grained aesthetic discrimination capabilities. The visualizations showcase test series from Natural, AIGC, and Cropping categories, where FGAesQ consistently produces accurate rankings that closely align with human aesthetic judgments. Compared to state-of-art IAA methods (Charm [3], MUSIQ [17]), our approach exhibits superior sensitivity to subtle aesthetic variations, including lighting conditions, color harmony, and compositional balance. This robust performance across different evaluation scenarios highlights FGAesQ's versatility and practical value for real-world applications, such as photo album management and curation, text-to-image generation optimization, and automated composition refinement for photography enhancement.

## 11. Limitations

**Dependency on Human Annotations.** The reliance on extensive human annotations for fine-grained aesthetic discrimination poses scalability challenges and introduces potential subjective biases. Developing automated or semi-automated annotation strategies that maintain annotation quality while reducing human effort remains an important direction for future work.

**Interpretability and Feedbacks.** While FGAesQ achieves accurate fine-grained aesthetic discrimination, generating specific and actionable feedback to explain its predictions remains challenging. Developing methods that articulate precise aesthetic rationales, such as identifying compositional flaws or suggesting targeted improvements, would significantly enhance the practical applicability of fine-grained IAA models. This promising yet challenging direction merits further exploration.

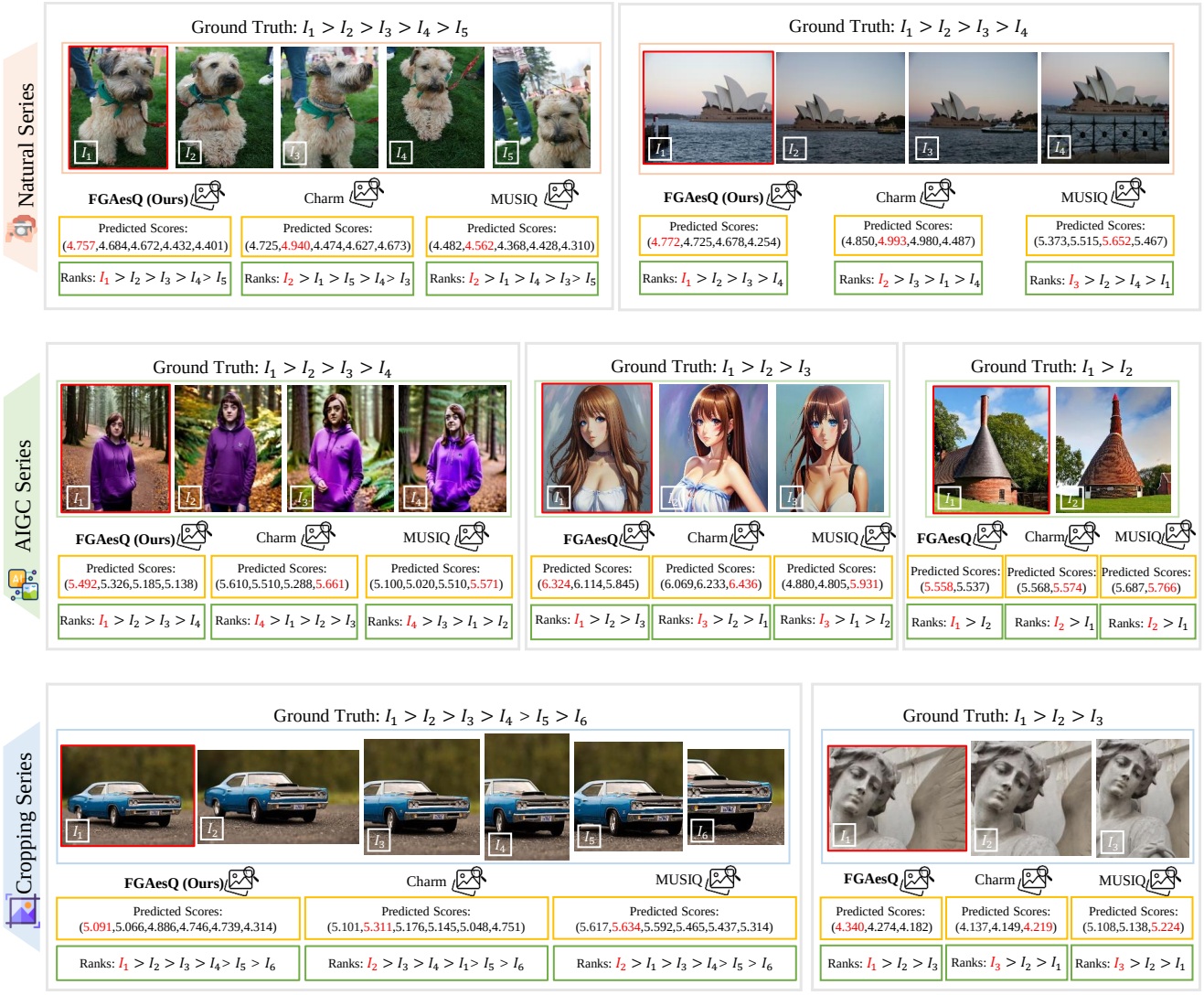



Figure 11. Additional visualization examples of FGAesQ on test series from Natural, AIGC, and Cropping categories.