*big data and cognitive computing*

MDPI

*Article*

# Unifying Aesthetic Evaluation via Multimodal Annotation and Fine-Grained Sentiment Analysis

**Kai Liu [1], Hangyu Xiong [2], Jinyi Zhang [3] and Min Peng [1,*]**

[1] School of Computer Science, Wuhan University, Wuhan 430072, China; zerkvii@whu.edu.cn
[2] Department of Computer Science, Technical University of Denmark (DTU), 2800 Copenhagen, Denmark; s250121@dtu.dk
[3] Department of Computer Science, University of California, Los Angeles, CA 90095, USA; jerryz12@g.ucla.edu
* Correspondence: pengm@whu.edu.cn

**Abstract**

With the rapid growth of visual content, automated aesthetic evaluation has become increasingly important. However, existing research faces three key challenges: (1) the absence of datasets combining Image Aesthetic Assessment (IAA) scores and Image Aesthetic Captioning (IAC) descriptions; (2) limited integration of quantitative scores and qualitative text, hindering comprehensive modeling; (3) the subjective nature of aesthetics, which complicates consistent fine-grained evaluation. To tackle these issues, we propose a unified multimodal framework. To address the lack of data, we develop the Textual Aesthetic Sentiment Labeling Pipeline (TASLP) for automatic annotation and construct the Reddit Multimodal Sentiment Dataset (RMSD) with paired IAA and IAC labels. To improve annotation integration, we introduce the Aesthetic Category Sentiment Analysis (ACSA) task, which models fine-grained aesthetic attributes across modalities. To handle subjectivity, we design two models—LAGA for IAA and ACSFM for IAC—that leverage ACSA features to enhance consistency and interpretability. Experiments on RMSD and public benchmarks show that our approach alleviates data limitations and delivers competitive performance, highlighting the effectiveness of fine-grained sentiment modeling and multimodal learning in aesthetic evaluation.

**Keywords:** aesthetic assessment; image caption; image sentiment; text sentiment; deep learning

Check for updates

Academic Editor: Giuseppe Maria Luigi Sarnè

Received: 17 November 2025
Revised: 13 January 2026
Accepted: 15 January 2026
Published: 22 January 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

The proliferation of social media platforms has fundamentally transformed multimedia content creation and consumption patterns. While democratizing information access, this expansion introduces significant challenges for automated quality assessment [1], particularly in aesthetic evaluation. As illustrated in Figure 1, user-generated content often contains fine-grained aesthetic feedback spanning multiple attributes, accompanied by subjective sentiment expressions, which further complicates reliable automatic assessment. Traditional manual filtering approaches remain prohibitively expensive and unscalable [2], necessitating robust computational solutions for content curation and AI-driven visual generation.

Computational aesthetic assessment aims to model human perceptions of visual beauty, with applications spanning image enhancement, content recommendation, and personalized media management [3]. This domain encompasses two primary tasks: Image Aesthetic Assessment (IAA), predicting numerical quality scores, and Image Aesthetic

Captioning (IAC), generating natural-language aesthetic descriptions. Despite substantial progress, existing methods suffer from fundamental limitations in dataset construction and modeling capability.

First, large-scale high-quality annotation remains a critical bottleneck. Benchmark datasets including AVA [4], AVA-comments [5], and PCCD [6] rely predominantly on point-based ratings or rule-driven heuristics. While efforts such as PIAA [3] improve annotation quality, scalability remains constrained by limited human resources, annotation inconsistency, and copyright restrictions. Recent work addresses ambiguity by redefining aesthetic ground truths beyond simple score aggregation [7], yet comprehensive solutions remain elusive.

Second, aesthetic perception inherently involves nuanced attributes—color harmony, compositional balance, emotional resonance—that conventional models struggle to capture from aggregated scores alone [8]. Many captioning approaches generate generic descriptions lacking attribute-specific evaluative depth. Recent research increasingly incorporates textual information as complementary modality. UFE-ISM integrates user traits and eye-tracking signals [9], while MIGF-Net models image-comment interactions through dual-stream fusion [10]. AMH-Net leverages large multimodal models for attribute-rich artistic descriptions [11]. Next-generation vision-language models such as LLaMA-Vision, LLaVA-NeXT [12], and InternVL 2 [13] demonstrate remarkable capabilities through large-scale pretraining. However, their application to fine-grained aesthetic evaluation—particularly structured attribute analysis and sentiment-aware modeling—remains underexplored. Our work specifically addresses aesthetic evaluation through systematic attribute decomposition and sentiment integration, complementing these general-purpose architectures.

To unify numerical precision with qualitative richness, we propose a tri-modal framework integrating: (1) quantitative IAA scores for objective measurement, (2) fine-grained attribute-level sentiment polarities capturing multi-dimensional aesthetic perception across color, composition, focus, and contrast, and (3) qualitative IAC descriptions articulating judgments in natural language. This holistic representation enables comprehensive aesthetic modeling by synthesizing numerical precision, sentiment granularity, and textual expressiveness.

We address data scarcity and modeling limitations through a comprehensive multimodal framework. Our contributions are threefold:

- **TASLP Pipeline and RMSD Dataset Construction.** Leveraging large language models, TASLP overcomes manual labeling limitations while maintaining annotation quality. The pipeline processes r/photocritique community content, converting user comments into structured evaluations across four dimensions: Color & Light, Composition, Focus, and Contrast. Each pipeline stage undergoes rigorous validation. RMSD comprises 61,224 images with 141,452 comments, averaging 2.3 comments per image, substantially exceeding existing dataset density while capturing diverse photographic styles from an authentic community.
- **LAGA: Graph-Based Aesthetic Scoring.** LAGA explicitly models attribute interdependencies via hierarchical graph convolutional networks. Two directed graphs capture category co-occurrence and category-sentiment dependencies, enabling nuanced scoring. This design enhances both performance and interpretability by revealing how aesthetic dimensions jointly contribute to assessment.
- **ACSFM: Attribute-Specific Caption Generation.** ACSFM integrates aesthetic category representations and LAGA-derived feedback into caption generation, producing detailed attribute-specific critiques beyond generic descriptions. Human and automatic evaluations confirm consistent superiority over strong baselines while maintaining competitiveness with large-scale models.

https://doi.org/10.3390/bdcc10010037

The paper proceeds as follows: Section 2 surveys related literature, Section 3 formalizes task definitions, Section 4 details methodology including TASLP, LAGA, and ACSFM, Section 5 presents experimental results, Section 6 provides qualitative analysis, and Section 7 concludes with future directions.

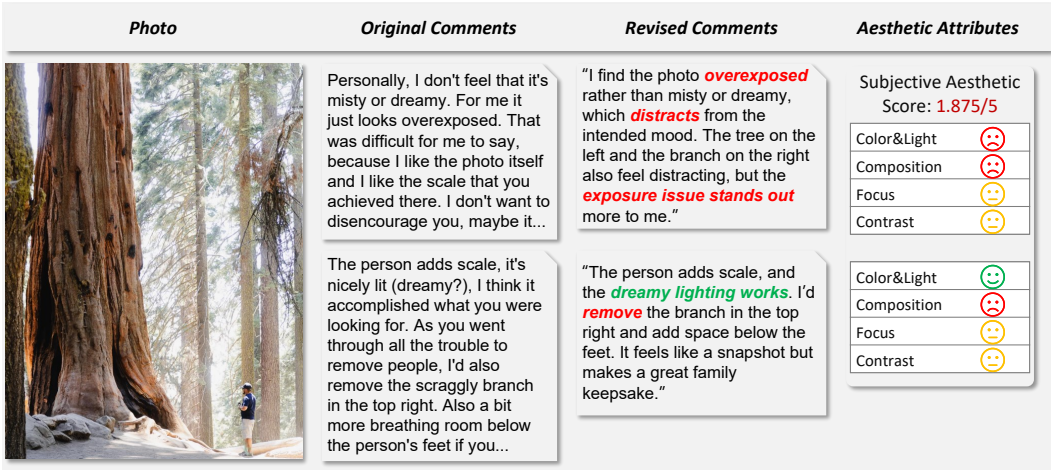


**Figure 1.** Example of a sample post with sentiment annotations across four aesthetic attributes: Color & Light, Composition, Focus, and Contrast. The figure illustrates the original comment, the revised version, and the corresponding aesthetic evaluation scores for each attribute.

## 2. Related Work

### *2.1. Image Aesthetic Assessment*

Image Aesthetic Assessment has evolved from distortion-based models to machine learning approaches utilizing support vector machines [14], subsequently shifting toward aesthetic value modeling [15]. Photo.NET [16] catalyzed datasets including AVA [4], PCCD [6], DPC-captions [17], and RPCD [18], though these suffer from limited annotations (AVA), single-annotator bias (PCCD), and data quality issues (RPCD).

Deep learning enabled CNN-based architectures such as RAPID [19] and Grad-CAM-enhanced ResNet-18. Subsequent datasets—AADB [20], CADB [21], PIAA [3]—incorporated manual attribute annotations but faced annotator inconsistency and limited community insights.

Recent research emphasizes multimodal integration. EHI-Net captured hierarchical modality interactions despite complexity and parameter sensitivity [22]. CDMC-Net enhanced cross-modal memory but encountered fusion inconsistencies and computational costs [23]. MIGF-Net explicitly models image-comment interplay with dedicated missing-comment handling [10]. AMH-Net targets artistic images via attribute-guided approaches using large language models for rich seven-dimensional descriptions [11]. Next-generation models like LLaMA-Vision, LLaVA-NeXT [12], and InternVL 2 [13] demonstrate remarkable vision-language understanding through large-scale pretraining. However, their application to fine-grained aesthetic evaluation—particularly attribute-level sentiment and detailed evaluative feedback—remains limited. Our work complements these general-purpose architectures by specifically addressing aesthetic-specific modeling through structured attribute analysis. These advances highlight a trajectory toward sophisticated multimodal fusion while underscoring persistent scalability and robustness challenges.

### *2.2. Image Aesthetic Captioning*

Paralleling assessment advances, image aesthetic captioning enriches evaluation by pairing images with detailed descriptions. Early approaches employed Bag-of-Words models and Fisher Vectors. Initial models AF and AO [6] generated captions based primarily

on overall quality. AMAN [24] focused on specific attributes—color harmony, composition, lighting—for nuanced descriptions.

Recent models BLIP [25] and VILA [26] advance the field further. BLIP leverages contrastive learning and CLIP [27] representations for high-fidelity captions, while VILA incorporates augmented samples for diversity. However, BLIP lacks explicit aesthetic supervision and VILA's augmented data can compromise aesthetic relevance. Multimodal large language models (MLLMs) such as LLaVA [28], Qwen-VL [29], and Pixtral demonstrate impressive general captioning capabilities. LLaVA connects CLIP's visual encoder with Vicuna through trainable projection. Qwen-VL incorporates position-aware vision-language alignment for spatial understanding. Pixtral employs efficient visual token processing for high-resolution inputs. While excelling at general description, these models lack optimization for aesthetic evaluation requiring explicit attribute modeling, sentiment polarities, and evaluative feedback. ACSFM addresses this gap through aesthetic-specific architectural components and training objectives. These limitations underscore the need to distinguish general captioning from aesthetic-specific generation.

## 3. Task Definition

Traditional aesthetic evaluation is formulated as:

$$v = \sum_{i=1}^{n} \mathrm{F}_i(I, \theta), \tag{1}$$

where $v$ denotes aesthetic score, $n$ represents feature count, $I$ is the input image ($H \times W \times 3$), $\theta$ denotes IAA model parameters, and $\mathrm{F}_i$ evaluates the $i$-th feature.

Integrating textual commentary yields Aesthetic Category Sentiment Analysis (ACSA). Given textual input $\{t_1, \ldots, t_k\}$, the objective extracts $\ell$ aesthetic categories $\{c_1, \ldots, c_\ell\}$ ($\ell = 4$) with sentiment-polarity pairs $\{(c_1, s_1), \ldots, (c_\ell, s_\ell)\}$. Aesthetic score computation:

$$\text{Score} = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{4} \left( c_\ell^i \cdot w_\ell^i \cdot s_\ell^i \right), \tag{2}$$

where $n$ represents comment count, and $\mathbf{w} = [w_1, \ldots, w_\ell]$ defines category weights.

Aesthetic comment generation parallels image captioning:

$$t_i = \mathrm{G}(\mathrm{p}(t_i \mid t_1, \ldots, t_{i-1}, I, \theta_c)), \tag{3}$$

where $t_i$ denotes the $i$-th word, p represents conditional probability, G is the decoding function, and $\theta_c$ denotes captioning model parameters.

## 4. Methods

Our approach comprises three components: TASLP pipeline leveraging large language models to construct RMSD from approximately 100,000 text-image pairs; LAGA model for IAA guided by ACSA; and ACSFM model for aesthetic captioning optimized using LAGA-derived feedback.

### 4.1. Textual Aesthetic Sentiment Labeling Pipeline

To comprehensively capture aesthetic information while minimizing manual effort, we propose TASLP (Figure 2), enhancing data representation through four stages: (1) aesthetic data acquisition and cleaning, (2) comment rewriting, (3) category clustering, and (4) category sentiment labeling. Each stage ensures high-quality, community-driven multimodal dataset construction.

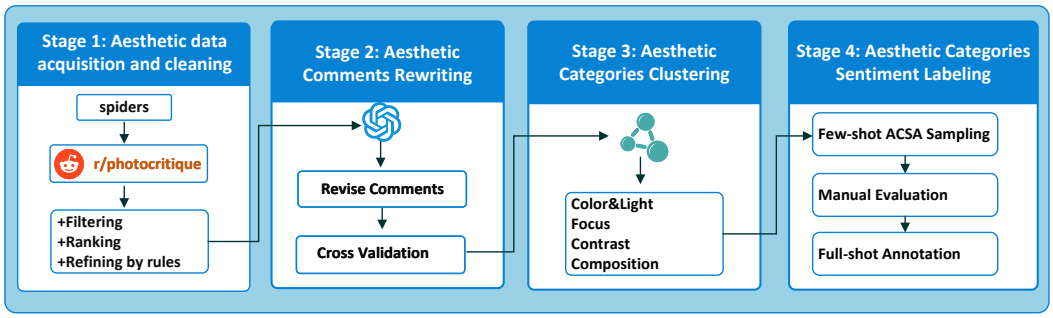


**Figure 2.** TASLP pipeline overview comprising four stages: (1) data acquisition and cleaning, (2) aesthetic comment rewriting, (3) aesthetic category clustering, and (4) aesthetic category sentiment labeling.

#### 4.1.1. Data Acquisition and Cleaning

Within Creative Commons framework, we accessed r/photocritique posts via Reddit's API, collecting 61,224 posts from 29,663 users and 141,452 comments from 33,541 users. We retained single-image posts, excluding multi-image submissions and comments with multimodal attachments.

Targeted filtering removed comments referencing fewer than three aesthetic attributes. For qualified comments, we analyzed distributions (Figure 3), sorting by Critique Points (https://www.reddit.com/r/photocritique/wiki/critiquepoints/, accessed on 15 October 2025) and retaining the top six when counts exceeded six.

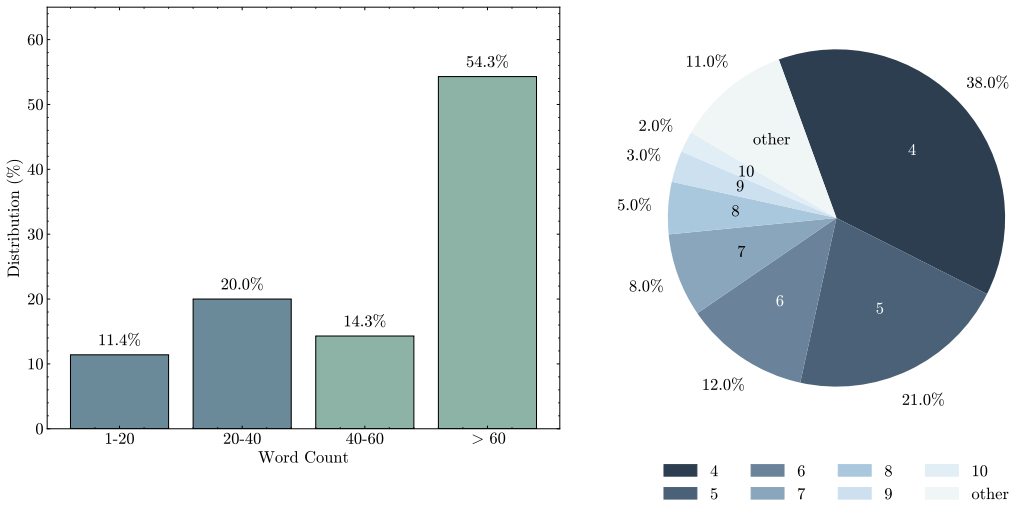


**Figure 3.** RMSD data distribution. The figure shows the distribution of word counts in user comments and the proportion of posts containing four or more comments.

Comments containing harmful language were removed. Data cleaning operations included HTML tag removal using regular expressions, emoji replacement using the Python Emoji library (version 2.10.1), and spelling correction using the TextBlob library (version 0.18.0).

#### 4.1.2. Dataset Biases and Limitations

The construction of RMSD introduces several inherent biases that warrant explicit acknowledgment. Source dependency bias stems from the exclusive reliance on r/photocritique, which may overrepresent Western aesthetic norms. Stylistic bias arises from the prevalence of critique-seeking images, leading to a skew toward medium-quality photographs. Demographic bias reflects the user composition of Reddit in terms of age, geography, and socioeconomic background. Temporal bias emerges from evolving aesthetic

preferences that may not be fully captured within the data collection period, while linguistic bias results from the exclusive focus on English-language content. These limitations should be carefully considered in practical deployment. Future work will explore broader data sources, diverse cultural contexts, extended timeframes, and multilingual settings.

#### 4.1.3. Aesthetic Comment Rewriting

Analysis identified two issues: excessively long sentences (Figure 3) exceeding model input limits, and colloquial expressions. We employed large language models (LLMs) for rewriting [30], with detailed prompts in Table 1.

**Table 1.** Model evaluation and comment revision prompts. The table presents a unified prompt used for model-based evaluation and aesthetic comment revision.

| **Unified Evaluation Prompt** |
|---|
| Part I: Model Cross Evaluation (Text Rewriting)<br>You are a professional text optimization assistant specializing in photography reviews. Rewrite the following review according to these requirements:<br>Preserve Sentiment: Maintain the original emotional tone (positive, negative, or neutral).<br>Retain Semantics: Preserve the core message and key points without introducing irrelevant content.<br>Condense Word Count: Remove redundancy while maintaining clarity and impact.<br>Enhance Expression: Improve fluency, coherence, and professionalism using precise language that highlights photographic elements.<br>Part II: Aesthetic Comment Revision (Quality Scoring) |
| You are an expert text evaluator assessing the quality of rewritten reviews. Score both the original and revised versions according to the following criteria:<br>Content Retention (Ci): 1—Significant loss or inconsistencies; 2—Partial loss or minor semantic drift; 3—Accurate retention of original content.<br>Fluency (Fi): 1—Frequent grammatical errors or awkward phrasing; 2—Mostly correct but somewhat unnatural; 3—Natural and smooth expression.<br>Readability (Ri): 1—Unclear structure or logic; 2—Generally clear but overly complex; 3—Clear, concise, and well-structured. |

We compared LLaMA-2 [31], Gemini [32], Qwen [33], and GPT-4o [34], evaluating content retention, fluency, and readability on three-level scales. Cross-evaluation yielded overall scores:

$$S_\phi = w_C \cdot C_i + w_F \cdot F_i + w_R \cdot R_i, \tag{4}$$

$$w_C = 0.5, w_F = 0.3, w_R = 0.2, \tag{5}$$

where $C_i, F_i, R_i$ denote content, fluency, readability scores (1–3 range), with empirically determined weights. Table 2 shows GPT-4o performed best, selected as core rewriting model.

**Table 2.** Cross-evaluation scores for comment revision models. Best scores in bold, second-best underlined. Final row shows column sums (200-entry sample).

| **Model** | **LLaMA-2-14B** | **Gemini** | **Qwen-14B** | **GPT-4o** |
|---|---|---|---|---|
| LLaMA-2-14B | <u>0.26</u> | 0.22 | 0.21 | **0.31** |
| Gemini | 0.23 | **0.34** | 0.26 | <u>0.32</u> |
| Qwen-14B | 0.20 | 0.28 | <u>0.36</u> | **0.37** |
| GPT-4o | 0.15 | 0.17 | <u>0.24</u> | **0.41** |
| Column Sum | **0.84** | **1.01** | **1.07** | **1.41** |

#### 4.1.4. LLM Annotation: Risks and Controls

LLM use introduces critical risks: linguistic homogenization potentially reducing stylistic variety, mitigated by prompts (Table 1) explicitly preserving sentiment and semantics

while addressing only length and clarity; artificial pattern injection introducing systematic biases, controlled through cross-model evaluation (Table 2) selecting GPT-4o for superior content retention (0.5 weight), ensuring minimal semantic drift; cultural bias amplification from overrepresented training perspectives, partially addressed via manual validation with multiple annotators (Fleiss' kappa: 0.78–0.83), though broader cultural validation remains necessary. Quality controls include: (1) retaining originals alongside rewritten versions, (2) filtering comments with fewer than three attributes, and (3) implementing multi-model cross-validation. Despite these measures, complete LLM bias elimination is impossible; future work should explore hybrid human-AI annotation approaches.

#### 4.1.5. Aesthetic Category Clustering

To systematically analyze aesthetic categories, we employ BERTopic [35] to perform semantic similarity-based clustering. BERTopic has demonstrated effectiveness across diverse domains [36,37]. The resulting clusters (Table 3) reveal four dominant aesthetic categories: Color & Light, Composition, Focus, and Contrast. This data-driven categorization establishes a principled framework for organizing heterogeneous aesthetic feedback, thereby supporting subsequent attribute-level sentiment analysis and model development through fine-grained, interpretable, and semantically grounded evaluation.

#### 4.1.6. Four-Attribute Framework Rationale

The four aesthetic attributes—Color & Light, Composition, Focus, and Contrast—derive from empirical clustering analysis (Table 3) and established photographic critique principles. Top-10 topics naturally group into: (1) Color & Lightencompassing illumination and color theory (lights, HDR, sky); (2) Composition capturing structural elements (composition, crop, foreground); (3) Focus addressing sharpness and depth-of-field (focus, subject); and (4) Contrast pertaining to tonal and dynamic range. This framework balances comprehensiveness with tractability, covering essential critique dimensions while remaining computationally manageable for LAGA's graph convolutional architecture. Alternative schemes with more categories (e.g., 7 in PCCD, 14 in AVA) were rejected due to increased annotation complexity without proportional performance gains and data sparsity in fine-grained categories, while fewer attributes would sacrifice modeling inter-attribute dependencies central to our graph-based approach.

**Table 3.** Top 10 topics from BERTopic clustering. Subjective attribute words in bold.

| Topic | Count |
|---|---|
| shot_subject_**focus**_bit | 0.332 |
| trees_tree_branches_**foreground** | 0.119 |
| sky_clouds_ground_cloud | 0.107 |
| bird_birds_feathers_duck | 0.081 |
| car_cars_**lights**_headlights | 0.075 |
| **hdr**_range_halo_effect | 0.063 |
| **composition**_critique_opinion_**composer** | 0.043 |
| dog_dogs_pup_doggo | 0.039 |
| **crop_cropping**_ratio_aspect | 0.036 |
| sky_**crop**_ratio_horizon | 0.035 |

#### 4.1.7. Aesthetic Category Sentiment Labeling

To facilitate aesthetic scoring, we annotate sentiment polarity through three steps:

1. Few-shot ACSA Sampling: Existing ABSA models require annotated datasets, limiting generalization. Prior research [38] demonstrates large language models achieve strong zero-shot performance when appropriately prompted. We utilized various prompts and models (Table 4) for ACSA sample generation.

**Table 4.** Prompts for Aesthetic Category Sentiment Analysis.

| | **Prompts for Aesthetic Category Sentiment Analysis** |
|---|---|
| Prompt I | |
| 1 | Extract paragraph attitude toward following dimensions. Attitude range: (positive, neutral, negative). |
| 2 | Extract paragraph attitude toward following dimensions. Attitude range: (positive, neutral, negative). If unrelated to dimension, attitude is neutral. |
| 3 | Provide sentiment polarity for specified dimensions only. Polarity range: (positive, neutral, negative). If unrelated, polarity is neutral. |
| 4 | Comment on paragraph sentiment polarity for specified dimensions. Provide polarity within (positive, neutral, negative). If unrelated, polarity is neutral. |
| Prompt II | |
| 1 | Dimensions: 1. color & light; 2. composition; 3. contrast; 4. focus. Paragraph: [Comment] |
| 2 | Dimensions: 1. Impression of color and light; 2. Impression of composition; 3. Impression of contrast; 4. Impression of focus. Paragraph: [Comment] |

Note: [Comment] represents aesthetic comments.

2. Manual Evaluation: Three graduate students with CET-6 certification performed sentiment annotation as ground truth. Fleiss' kappa coefficients (0.83, 0.78, 0.81) indicate strong inter-annotator consistency. Figure 4 shows GPT-4o achieved highest accuracy (85.8% peak in Prompts $I_4$ and $II_2$), consistently outperforming alternatives.

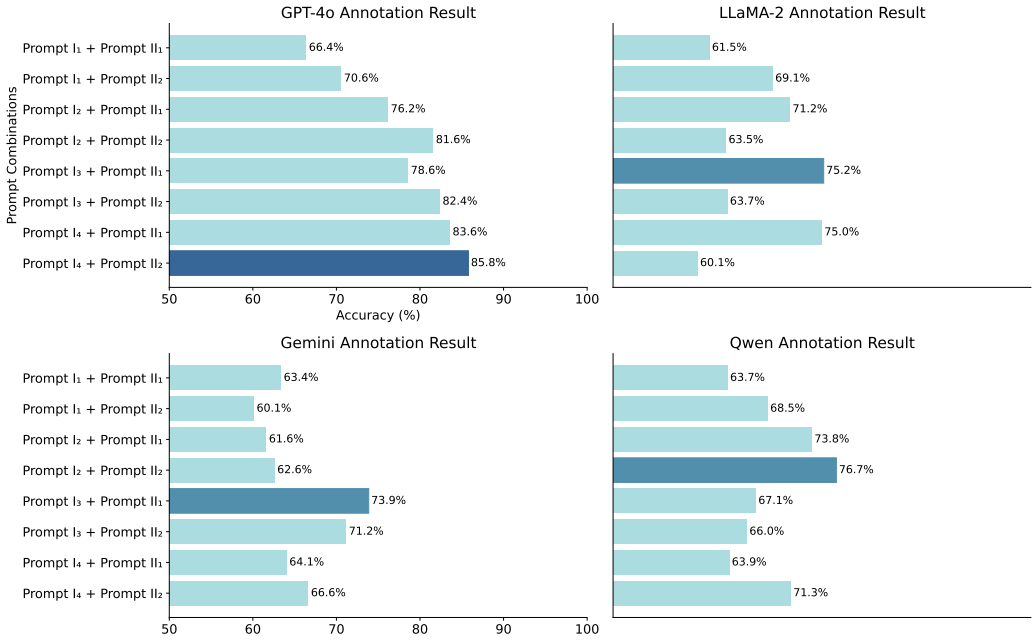


**Figure 4.** Performance comparison across prompt configurations for aesthetic sentiment labeling.

3. Full-shot Annotation: GPT-4o comprehensively annotated RMSD, yielding rich text-based and score-based annotations. Table 5 presents detailed filtering statistics. From initial 61,224 posts with 141,452 comments, we excluded 8247 multi-image posts (13.5%), removed 23,156 comments with fewer than three attributes (16.4%), filtered 4893 comments with multimodal attachments (3.5%), and eliminated 2,318 comments with harmful language (1.6%). Final dataset retains 53,977 posts with 110,085 comments, representing 88.2% post retention and 77.8% comment retention, demonstrating rigorous quality control while maintaining substantial coverage.

**Table 5.** Filtering Statistics for RMSD Dataset Construction.

| Filtering Criterion | Excluded | Percentage | Remaining |
|---|---|---|---|
| Initial collection | - | - | 61,224 posts, 141,452 comments |
| Multi-image posts | 8247 | 13.5% | 52,977 posts |
| Comments with <3 attributes | 23,156 | 16.4% | 118,296 comments |
| Multimodal attachments | 4893 | 3.5% | 113,403 comments |
| Harmful language | 2318 | 1.6% | 111,085 comments |
| Final dataset | - | - | 53,977 posts, 111,085 comments |

#### 4.1.8. Dataset Split

RMSD is partitioned into training, validation, and test sets with ratio 8:1:1, yielding 43,182 posts for training, 5397 for validation, and 5398 for testing. Splitting is performed at post level to prevent data leakage from comments belonging to the same image appearing across splits. Category distribution remains consistent across splits (verified via chi-square test, $p > 0.05$), ensuring representative evaluation.

Table 6 compares RMSD with contemporary benchmarks, highlighting differences in size, scale, categories, attribute diversity, and annotation types.

**Table 6.** Summary of publicly available image aesthetic assessment datasets.

| Dataset | Images | Scale | Topics | Attributes | Comments | Annotations |
|---|---|---|---|---|---|---|
| Photo.Net | 20 K | 1–7 | - | - | - | Score |
| CUHK-PQ | 17 K | 0–1 | 7 | - | - | Label |
| AVA | 250 K | 1–10 | 66 | 14 | - | Score |
| AVA-Comments | 250 K | 1–10 | 66 | - | 1.5 M | Text |
| PCCD | 4 K | 1–10 | - | 7 | 29 K | Text |
| DPC-Captions | 110 K | 1–10 | - | 5 | 2.4 M | Text |
| RPCD | 70 K | 1–10 | - | 5 | 219 K | Score |
| RMSD | 61 K | 1–5 | - | 4 | 141 K | Text & Score |

### *4.2. Layered Aesthetic Graph Convolutional Assessment Network*

To better leverage aesthetic category characteristics, we propose a novel scoring approach emphasizing textual features, as comments represent highly distilled aesthetic information. Given an image with comment set (minimum four), our approach calculates scores solely from textual information. This text-centric design is motivated by two observations: (1) user comments in RMSD already encapsulate visual aesthetic judgments in linguistic form, providing richer semantic signals than raw pixels; (2) ablation studies (Section 5.5) confirm text-based features substantially outperform image-only baselines, validating this architectural choice. Future work may explore hybrid text-image scoring for scenarios with limited textual feedback.

The framework (Figure 5) employs BERT [39] as foundational encoder capturing semantic content and positional encodings. To enhance interpretability, we incorporate sentiment information from SentiWordNet lexicon. The Encoding Module produces:

$$e_s = \text{Bert}(e_{ctx} + e_{pos} + e_{sentiment}) \tag{6}$$

where $e_s$ represents total embeddings, $e_{ctx}$, $e_{pos}$, $e_{sentiment}$ denote contextual, positional, sentiment embeddings.

To model connections between sentiment categories and sentiment polarity, we define two directed graphs capturing relationships between aesthetic categories and corresponding polarities. These structures explicitly represent statistical co-occurrence patterns: category adjacency matrix $A^c$ captures how frequently different attributes (e.g., color and composition) are discussed together, while sentiment adjacency matrix $A^s$ encodes condi-

tional probability of sentiment polarity for one category given another. This graph-based representation enables leveraging inter-attribute dependencies—for instance, positive composition evaluation may influence expected focus sentiment, as these attributes often relate in photographic critique. Relationships are represented via adjacency matrices:

$$A^c_{i,j} = \begin{cases} \frac{\text{count}(c_i, c_j)}{\text{count}(c_i)+1}, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}, \tag{7}$$

$$A^s_{i,j} = \begin{cases} \frac{\text{count}((s_j|c_j), c_i)}{\text{count}(c_i)+1}, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}, \tag{8}$$

where $\text{count}(c_i)$ denotes $i$-th category frequency, $\text{count}(c_i, c_j)$ represents co-occurrence frequency, and $\text{count}((s_j \mid c_j), c_i)$ specifies sentiment polarity $s_j$ frequency for $j$-th category given $i$-th category presence.

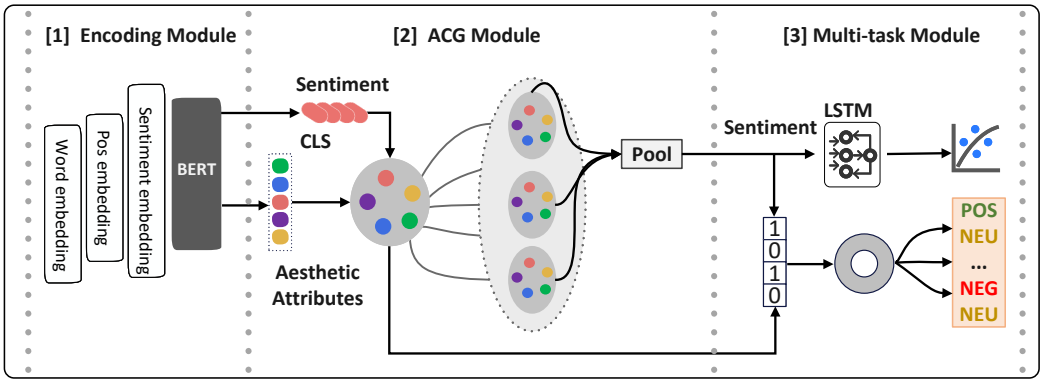


**Figure 5.** LAGA framework architecture. The model consists of three main components: (1) Encoding Module, which extracts semantic, positional, and sentiment-aware embeddings using BERT; (2) Aesthetic Category Group (ACG) Module, which aggregates and pools representations of different aesthetic attribute categories; (3) Multi-task Module, which jointly performs aesthetic category sentiment analysis (ACSA) and image aesthetic assessment (IAA). In the figure, different colored nodes represent embeddings associated with different aesthetic attribute categories, including Color & Light, Composition, Focus, and Contrast. CLS denotes the classification token, Enc represents the encoder module, $\{v_1, v_2, v_3, v_4\}$ correspond to the four aesthetic category representations, and the binary vector $(1, 0, 1, 0)$ indicates the presence (1) or absence (0) of the corresponding aesthetic categories.

We employ two-layer graph convolutional network (GCN) capturing category and category-sentiment relationships. GCN architecture performs message passing over constructed graphs, aggregating neighboring node information to update representations iteratively. Each GCN layer applies normalized graph convolution followed by non-linear activation, learning higher-order dependencies beyond direct edges. Parameters include: hidden dimension $d_h = 256$, GCN dropout rate $p_{gcn} = 0.3$, with two GCN layers as preliminary experiments showed diminishing returns beyond this depth. Normalization $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ ensures symmetric graph structure treatment and prevents gradient instability.

$$h_{\ell+1} = \text{GELU}\left(\tilde{D}_c^{-1/2}\tilde{A}^c\tilde{D}_c^{-1/2}h_\ell W_\ell^c + b_\ell^c\right), \tag{9}$$

$$\hat{F}_\ell = \text{TanH}(W_\ell^{cs}[h_{\ell+1} \oplus S_\ell] + b_\ell^{cs}), \tag{10}$$

$$\tilde{F}_\ell^s = \text{ReLU}\left(\tilde{D}_s^{-1/2}\tilde{A}^s\tilde{D}_s^{-1/2}\hat{F}_\ell W_\ell^s\right), \tag{11}$$

where $W$, $b$ represent graph convolution weights and biases, $h_\ell$ denotes $\ell$-th layer hidden state, $S_\ell$ represents $\ell$-th category sentiment polarity, and $\oplus$ denotes concatenation.

We obtain $\ell + 1$-th category sentiment polarity distribution via pooling:

$$S_{\ell+1} = \text{MaxPool}(\tilde{F_\ell}^{pos}, \tilde{F_\ell}^{neu}, \tilde{F_\ell}^{neg}). \tag{12}$$

Subsequently, we derive category $\text{p}(c_i)$, sentiment $\text{p}(s_{i,j})$, and aesthetic score $\tilde{y}^s$ representations:

$$\text{p}(c_i) = \text{Sigmoid}(W_i^c h_i + b_i^c), \tag{13}$$

$$\text{p}(s_{i,j}) = \text{Softmax}(W^s S_{\ell+1} + b^s), \tag{14}$$

$$\hat{\text{y}}(i,j) = \text{Argmax}(\text{p}_i, \text{p}(i,j)), \tag{15}$$

$$\tilde{y}^s = \text{LSTM}(\tilde{F}_\ell^s), \tag{16}$$

where $W^s$, $b^s$ are shared across sentiment polarity computation.

We employ multi-task learning with three sub-tasks: aesthetic category recognition, sentiment polarity prediction, and aesthetic scoring. Loss functions:

$$\mathcal{L}_c = -\sum_{i=1}^{5} y_i^c \log \text{p}(c_i) + (1 - y_i^c) \log(1 - \text{p}(c_i)), \tag{17}$$

$$\mathcal{L}_s = -\sum_{i=1}^{5} \sum_{j=1}^{3} y_{i,j} \log \text{p}(s_{i,j}), \tag{18}$$

$$\mathcal{L}_a = \frac{1}{n} \sum_{k=1}^{n} (\tilde{y}_k^s - y_k)^2, \tag{19}$$

$$\mathcal{L}_t = \alpha \mathcal{L}_c + \beta \mathcal{L}_s + \gamma \mathcal{L}_a. \tag{20}$$

Loss weights $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.3$ were determined via grid search over ranges $\alpha \in [0.3, 0.5]$, $\beta \in [0.2, 0.4]$, $\gamma \in [0.2, 0.4]$ with step size 0.1 on 10% validation set. These values balance three objectives: category recognition requires slightly higher weight ($\alpha$) as foundational classification, while sentiment prediction ($\beta$) and scoring ($\gamma$) contribute equally to aesthetic assessment. Alternative configurations showed degradation: increasing $\gamma$ beyond 0.3 caused overfitting on score prediction compromising attribute-level interpretability, while decreasing $\alpha$ below 0.4 impaired category detection accuracy.

### *4.3. Aesthetic Category Sentiment Fused Multimodal Model*

To jointly perform IAA and IAC within unified framework, we propose Aesthetic Category Sentiment Fused MultiModal model (ACSFM) based on LAGA. ACSFM comprises: (1) Multi-Aesthetic Visual Channel Encoder (MACE), (2) Transformer-based Aesthetic Caption Decoder (TACD) [40], and (3) LAGA-based Aesthetic Category and Score Feedback Module. Figure 6 illustrates overall architecture.

#### 4.3.1. Multi-Aesthetic Visual Channel Encoder

Image encoder employs CNNs or patch-based methods like ViT [41]. Features:

$$e_i = \text{Enc}(I_a, I_r, I_g, I_b) = \{v_1, v_2, v_3, v_4\}, \tag{21}$$

$$h_I = \text{FFN}(e_1 \oplus e_2 \oplus e_3 \oplus e_4). \tag{22}$$

Image $I$ is single-channel vector where $I_a, I_r, I_g, I_b$ correspond to alpha, red, green, blue components. Feature vectors $e_i$ extract from corresponding channels, $\{v_1, v_2, v_3, v_4\}$ capturing individual representations. Concatenated features process via feedforward neural network (FFN) computing hidden state $h_I$, subsequently feeding TACD through cross-attention.

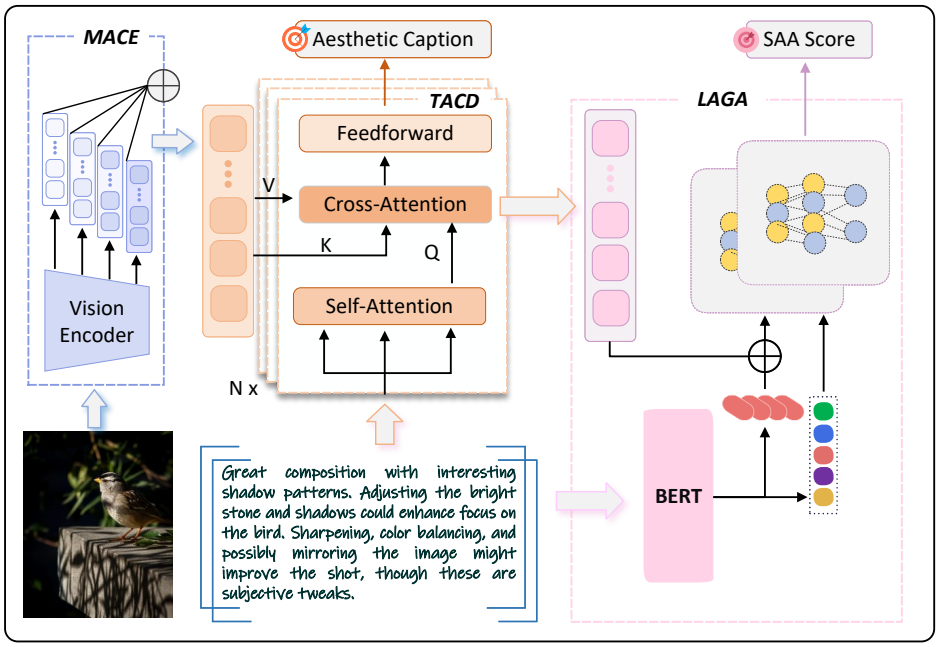


**Figure 6.** ACSFM model architecture for unified aesthetic evaluation. Red circles in the LAGA component represent sentiment polarity nodes (positive, neutral, negative), while colored circles in MACE indicate channel-specific visual feature representations.

#### 4.3.2. Transformer-Based Aesthetic Caption Decoder

Transformer [40] decoder widely demonstrates powerful text generation capability. In TACD, we integrate fused aesthetic visual category features from MACE with textual features:

$$e_T = \mathrm{Enc}^{\mathrm{T}}(e_{ctx}, e_{pos}, e_{sentiment}), \tag{23}$$

$$h_T = \mathrm{Softmax}(\frac{e_T^q \cdot e_T^k}{\sqrt{d_k^T}} \cdot e_T^v), \tag{24}$$

$$h_F = \mathrm{Softmax}(\frac{h_T^q \cdot h_I^k}{\sqrt{d_k^T}} \cdot h_I^v). \tag{25}$$

Here, $e_T$ encompasses textual embedding across multiple dimensions, $h_T$ denotes self-attention outcome, and $h_F$ characterizes cross-attention fusion outcome.

#### 4.3.3. LAGA Feedback Module

To enhance aesthetic caption generation, we integrate aesthetic categories and scores as feedback through LAGA. Sentiment polarity distribution concatenates with fused multimodal features from final Transformer decoder layer. Resulting features pass through linear layer:

$$e_s = \mathrm{Bert}(e_{ctx} + e_{pos} + e_{sentiment}) + W \cdot hf + b, \tag{26}$$

where $hf$ represents fused features.

Beyond LAGA, we employ multi-task learning jointly optimizing aesthetic generation and accuracy. Framework incorporates two loss functions:

$$\mathcal{L}_{gen} = -\frac{1}{N}\sum_{i=1}^{N} \log \mathrm{p}(t_i \mid t_{<i}, h_T), \tag{27}$$

$$\mathcal{L}_a = \frac{1}{n}\sum_{i=1}^{n} (\tilde{y}_i^s - y_i)^2, \tag{28}$$

$$\mathcal{L}_t = \alpha \mathcal{L}_{gen} + \beta \mathcal{L}_a, \tag{29}$$

where $\mathcal{L}_{gen}$ represents auto-regressive generation loss, $\mathcal{L}_a$ corresponds to IAA task loss. Coefficients $\alpha, \beta$ are empirically set to 0.3 and 0.5 respectively.

## 5. Experiments

### 5.1. Baselines

We compare against representative baselines across three categories: text-based, image-based, and multimodal models.

**Text-based models:**

- BART-Generation [42]: Sequence-to-sequence model leveraging generative learning for aspect category sentiment analysis.
- HGCN [43]: Hierarchical graph convolutional network modeling aspect term-sentiment dependencies.
- AAGCN [44]: Graph-based model incorporating external knowledge and beta-distribution weighting for aspect-level sentiment classification.

**Image-based models:**

- VGG-16 [45]: Classical CNN architecture for image quality and aesthetic assessment.
- ResNet-34 [46]: Residual network mitigating gradient degradation.
- ViT-Base [47]: Transformer-based vision model capturing global dependencies via patch-level representations.

**Multimodal models:**

- ViT-GPT2 [48]: Vision-language model combining ViT and GPT2.
- BLIP [25]: Contrastively pretrained vision-language framework for image-text tasks.
- LLaVA-1.5 [28]: Multimodal large language model connecting CLIP visual encoder with Vicuna-13B via instruction tuning.
- AestheticCLIP: Aesthetic-oriented CLIP variant trained on image-aesthetic text pairs for zero-shot assessment.

### 5.2. Implementation Details

Text-based models extend sequence lengths while retaining original training settings. Image-based models use standardized $224 \times 224$ inputs with adapted output heads, batch size 64, Adam optimizer (weight decay $4 \times 10^{-4}$). ViT employs patch size 16, trained 25 epochs with learning rate $3 \times 10^{-3}$, while ResNet-34 and VGG-16 train 30 epochs at $4 \times 10^{-4}$. LAGA hyperparameters follow HGCN baseline.

For aesthetic captioning, BLIP and ViT-GPT2 initialize from BLIP pretrained on COCO [49], fine-tuning 30 epochs with learning rate $5 \times 10^{-5}$ using 384-pixel inputs. LLaVA-1.5 uses public 13B checkpoint with CLIP ViT-L/14, fine-tuned 10 epochs (learning rate $2 \times 10^{-5}$, batch size 8, $336 \times 336$ inputs). AestheticCLIP adopts pretrained aesthetic checkpoint, adapting output layer for four-category sentiment classification, trained 20 epochs at $1 \times 10^{-4}$. All experiments conduct on 8 NVIDIA RTX 3090Ti GPUs.

### 5.3. Evaluation Metrics

For ACSA, we report Accuracy, Precision, Recall, and F1 score jointly evaluating correctness and false positive/negative balance. We note that Specificity (true negative rate) is less informative for our multi-class sentiment task where neutral predictions dominate; instead, per-class F1 scores capture class-specific performance more meaningfully.

For aesthetic scoring, we use PLCC and SRCC measuring linear and rank correlations between predictions and ground-truth. We report Two-Category Accuracy (2-Cate) for

high/low aesthetic discrimination and Difference Accuracy (D $\pm x$) evaluating tolerance-based continuous score prediction.

For aesthetic captioning, we adopt BLEU [50] and ROUGE-L [51]. BLEU (Bilingual Evaluation Understudy) assesses n-gram precision (BLEU-1 to BLEU-4), measuring how many n-grams in generated captions appear in references. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) measures longest common subsequence overlap, emphasizing content coverage and sequential coherence. Together, they provide complementary caption quality evaluation.

### 5.4. Results and Discussion

Not all baselines support ACSA, IAA, and IAC tasks, categorized into text-based, image-based, and multimodal models. Text-based models benchmark ACSA against LAGA, all models apply to IAA, only multimodal models handle IAC.

#### 5.4.1. ACSA Task on RMSD

Table 7 indicates HGCN and LAGA outperform BART-Generation and AAGCN in ACSA performance. GCN-based models' strong performance highlights value in multi-layer convolutional networks exploring connections between aesthetic categories and sentiment. BART-Generation's slightly lower performance suggests template similarity to pre-training templates enhances semantic knowledge usage but falls short of deeper understanding. AAGCN's performance indicates introducing noisy external knowledge can misguide training, influenced by quality and introduction methods.

**Table 7.** Text-based models and LAGA comparison on ACSA task. Best results in bold.

| Model | Acc. (%) | F1 | Precision | Recall |
|---|---|---|---|---|
| BART-Generation [42] | 82.77 | 76.73 | 75.60 | **77.91** |
| HGCN [43] | 82.89 | 76.94 | 77.64 | 76.25 |
| AAGCN [44] | 77.91 | 68.14 | 70.93 | 65.57 |
| LAGA (ours) | **83.02** | **77.05** | **78.52** | 75.64 |

Excluding LAGA, three baseline ACSA models approach ACSA from distinct perspectives: utilizing pre-trained language model knowledge, exploring deep relationships via graph convolutional networks, and introducing external knowledge bases. These methods achieve 77.9–82.9% accuracy and 68.1–76.9 F1 scores on RMSD, consistent with benchmark performance, confirming RMSD's feasibility for ACSA and indicating improvement potential. To assess statistical significance, we conducted McNemar's $\chi^2$ test on classification predictions. Comparing LAGA against HGCN (strongest baseline) yielded $\chi^2 = 12.43$ ($p < 0.001$), indicating LAGA's 0.13% accuracy improvement is statistically significant despite small margin. Similar tests against BART-Generation ($\chi^2 = 15.87$, $p < 0.001$) and AAGCN ($\chi^2 = 89.34$, $p < 0.001$) confirm LAGA's gains represent genuine modeling improvements rather than random chance.

#### 5.4.2. Aesthetic Scoring Assessment

Table 8 shows predicted aesthetic scores strongly correlate with ACSA outputs. Among text-based baselines, HGCN outperforms BART-Generation and AAGCN, achieving SRCC 0.832 and PLCC 0.831, demonstrating sentiment-aware modeling effectiveness for aesthetic scoring. This indicates aesthetic scores inferred from attribute-level sentiments are more interpretable and predictable than conventional holistic scores.

**Table 8.** Model performance comparison on IAA scoring task using RMSD dataset. Best results in bold. Statistical significance denoted by * ($p < 0.05$) and ** ($p < 0.01$) via paired $t$-tests against second-best model.

| Model | SRCC | PLCC | 2-Cate (%) | D ± 0.5 (%) |
|---|---|---|---|---|
| BART-Generation [42] | 0.805 | 0.807 | 82.13 | 72.57 |
| HGCN [43] | 0.832 | 0.831 | 83.98 | 71.03 |
| AAGCN [44] | 0.700 | 0.662 | 81.94 | 54.86 |
| VGG-16 [45] | 0.068 | 0.041 | 65.50 | 35.00 |
| ResNet-34 [46] | 0.096 | 0.070 | 65.00 | 37.50 |
| ViT-Base [41] | 0.132 | 0.093 | 69.50 | 38.00 |
| AestheticCLIP | 0.456 | 0.442 | 74.23 | 45.67 |
| ViT+GPT2 (+ACSFM ours) | 0.204 | 0.218 | 59.51 | 25.72 |
| BLIP (+ACSFM ours) | 0.615 | 0.610 | 73.16 | 37.84 |
| LAGA (ours) | **0.842** ** | **0.836** ** | **85.67** * | **73.19** ** |

Image-only models perform poorly. ViT-Base yields best results among ViT-Base, ResNet-34, VGG-16, achieving only 0.132 SRCC and 0.093 PLCC, highlighting difficulty in directly regressing images to aesthetic scores. AestheticCLIP attains moderate performance (SRCC: 0.456) but remains substantially inferior to LAGA, suggesting generic aesthetic representations insufficient without explicit category- and sentiment-level modeling.

Multimodal models consistently outperform image-only approaches. BLIP(+ACSFM) surpasses ViT-GPT2(+ACSFM), attributed to stronger text encoding capacity of BERT compared to GPT2. These results confirm aesthetic comments provide valuable guidance for learning aesthetic features from images, and joint image-text modeling suits aesthetic evaluation. Paired $t$-tests verify statistical significance of LAGA's gains over HGCN across SRCC ($p = 0.002$), PLCC ($p = 0.005$), and D ± 0.5 ($p < 0.001$), confirming robustness of improvements.

#### 5.4.3. Aesthetic Image Caption Generation

Table 9 shows BLIP and ViT-GPT2 exhibit superior aesthetic caption generation prowess. ViT-GPT2 consistently outperforms BLIP across all metrics, attributed to notably enhanced decoder. With ACSFM branch, models demonstrate substantial enhancement, consistently surpassing previous iterations. This observation confirms robust correlation between aesthetic scores and comments, affirming predictive efficacy of aesthetic score estimation. This predictive task complements caption generation, effectively augmenting models' capacity to assimilate image and comment features. LLaVA-1.5, despite being large-scale multimodal LLM, underperforms compared to specialized models on aesthetic captioning. This demonstrates general-purpose vision-language models require specific adaptation for attribute-focused evaluative caption generation, as training primarily targets descriptive rather than critical commentary. These outcomes demonstrate RMSD dataset aesthetic score reliability.

**Table 9.** Model performance comparison on aesthetic image captioning task using RMSD dataset. Best results in bold.

| Model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE |
|---|---|---|---|---|---|
| BLIP [25] | 41.65 | 17.03 | 6.70 | 2.71 | 20.27 |
| BLIP(+ACSFM ours) | **45.92** | **21.79** | **10.31** | **5.30** | **21.47** |
| ViT-GPT2 [48] | 51.39 | 24.78 | 10.88 | 4.95 | 23.54 |
| ViT-GPT2(+ACSFM ours) | **52.27** | **27.32** | **12.09** | **5.52** | **23.89** |
| LLaVA-1.5 | 38.42 | 15.67 | 6.23 | 2.84 | 19.35 |

### *5.5. Ablation Study*

#### 5.5.1. Module Function Analysis in LAGA for ACSA and IAA

To investigate each module's functionality in LAGA, we conducted ablation experiments on RMSD for both sentiment classification (ACSA) and image aesthetic assessment (IAA). We constructed three variants: w/o Sentiment (removes sentiment embedding and fusion modules), w/o ACG (uses entire sentence representation instead of aesthetic categories), and w/o A-BLSTM (replaces bidirectional LSTM with linear layer).

Results (Tables 10 and 11) reveal key findings: For ACSA, all components showed significant importance. Removing sentiment information decreased accuracy by 9.90 and F1 by 7.63 points. Removing A-BLSTM reduced accuracy by 8.70 and F1 by 7.33 points. Most critically, removing ACG caused largest drops: 11.92 points in accuracy and 9.22 points in F1, demonstrating its indispensable role in capturing contextual information for sentiment classification.

**Table 10.** LAGA variant performance comparison on ACSA task. Best results in bold, second-best underlined, decreases marked with downward arrow and decrease amount.

| Model | ACC | F1 | Precision | Recall |
|---|---|---|---|---|
| LAGA | **83.02** | **77.05** | **78.52** | **75.64** |
| w/o sentiment | 73.12↓(9.90) | <u>69.42</u>↓(7.63) | 69.18↓(9.34) | 69.67↓(5.97) |
| w/o ACG | 71.10↓(11.92) | 67.83↓(9.22) | 68.32↓(10.20) | 67.34↓(8.30) |
| w/o A-BLSTM | <u>74.32</u>↓(8.70) | 69.72↓(7.33) | <u>70.12</u>↓(8.40) | <u>69.32</u>↓(6.32) |

**Table 11.** LAGA variant performance comparison on IAA task. Best results in bold, second-best underlined, decreases marked with downward arrow and decrease amount.

| Model | SRCC | PLCC | 2-Cate (%) | D $\pm$ 0.5 (%) |
|---|---|---|---|---|
| LAGA | **0.842** | **0.836** | **85.67** | **73.19** |
| w/o sentiment | <u>0.833</u>↓(0.009) | <u>0.831</u>↓(0.005) | <u>85.05</u>↓(0.62) | 71.64↓(1.55) |
| w/o ACG | 0.812↓(0.030) | 0.808↓(0.028) | 83.36↓(2.31) | 69.65↓(3.54) |
| w/o A-BLSTM | 0.811↓(0.031) | 0.812↓(0.024) | 83.84↓(1.83) | <u>69.77</u>↓(3.42) |

For IAA, ACG and A-BLSTM were more critical. Removing sentiment had minimal impact (SRCC $-0.009$, PLCC $-0.005$). However, removing ACG or A-BLSTM significantly degraded performance (SRCC $-0.030$ and $-0.031$ respectively), with ACG removal causing notable 3.54 point drop in D $\pm$ 0.5 (%).

Overall, ACG serves as backbone for both tasks, providing essential contextual understanding, while sentiment module primarily benefits ACSA, and A-BLSTM refines predictions in both tasks.

#### 5.5.2. Module Function Analysis in ACSFM for IAC

To investigate each component's role in ACSFM for IAC task, we conduct module ablation experiments on RMSD, including removing aesthetic scoring module w/o LAGA replacing cross-modal fusion with direct concatenation w/o ACD, and using single encoder instead of multiple encoders w/o MACE. Experimental results in Figure 7 show aesthetic scoring module contributes most to performance, with BLEU-1 decreasing by 8.01 points (from 52.27 to 44.26) and CIDEr decreasing by 0.018 after removal, indicating aesthetic feedback plays core role in capturing aesthetic properties and optimizing text generation. Cross-modal fusion mechanism has second largest impact, with BLEU-2 and BLEU-3 decreasing by 8.0 and 5.8 points respectively after removal, demonstrating cross-attention mechanism outperforms simple concatenation in integrating multimodal information. Multi-encoder structure has relatively smaller impact, with BLEU-4 decreasing by only 1.5 points after removal; although single encoder captures key visual features, multiple

encoders' rich detail information further enhances model performance. Overall, component importance ranking: aesthetic scoring module > cross-modal fusion mechanism > multi-image encoder.

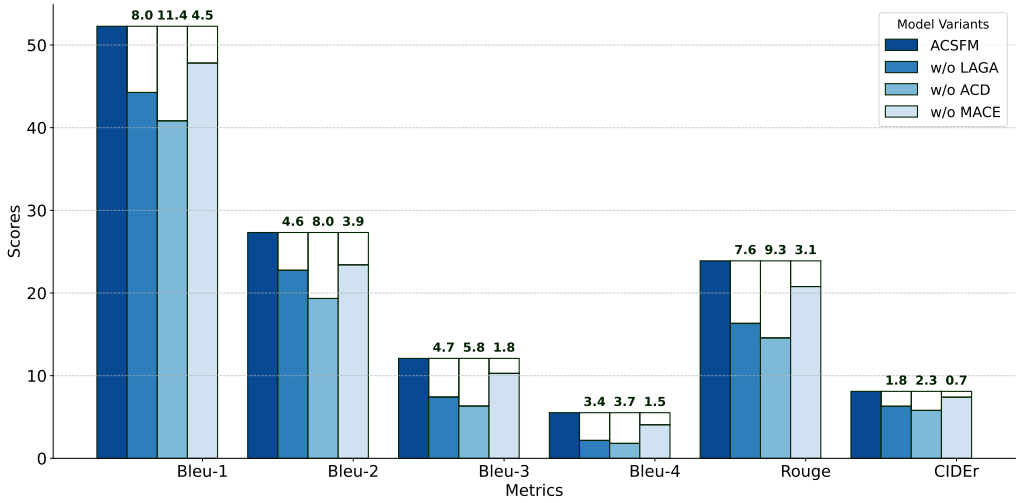


**Figure 7.** ACSFM variant performance comparison on IAC task.

### *5.6. Comparison to Large Language Models*

Large language models increasingly influence vision-language research. We evaluated LLaMa3-11B, InstructBlip-13B, and Qwen2-VL-7B on ACSA, IAA, and IAC tasks using 200 test samples. Figure 8 shows ACSFM consistently delivers superior performance, achieving highest scores across BLEU-1 (53.8), BLEU-2 (28.4), ROUGE (24.8), and CIDEr (8.5). In comparison, LLaMa3-11B-Vision_Instruct, nearest competitor, attains 51.2 for BLEU-1 and 7.4 for CIDEr but falls short. Smaller models like Qwen2-VL-7B-Instruct and InstructBLIP-7B exhibit significantly lower results. ACSFM's superior performance is attributed to sophisticated cross-modal feature integration and aesthetic scoring mechanisms, underscoring pivotal role of architectural design and model scale in capturing nuanced visual-textual relationships.

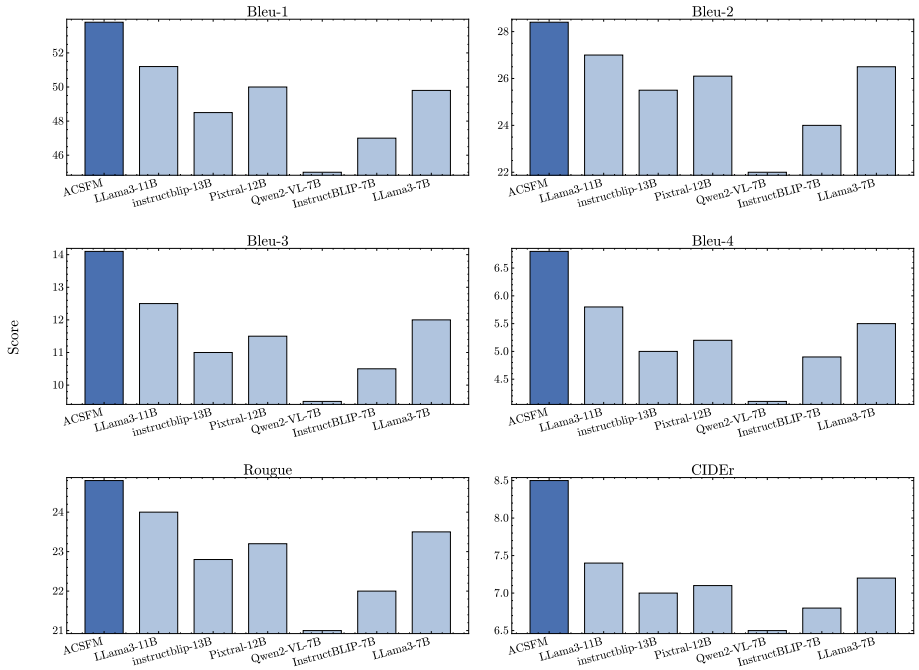


**Figure 8.** ACSFM performance comparison against large language models on aesthetic evaluation tasks. BLEU metrics follow standard definitions [50].

### *5.7. Human Evaluation*

To complement automatic metrics, we conducted human evaluation of aesthetic captioning on randomly selected 100 test samples. Three expert evaluators (graduate students with photography backgrounds and CET-6 certification) assessed generated captions across three dimensions: (1) Aesthetic Relevance (1–5 scale): whether caption addresses relevant aesthetic attributes; (2) Evaluative Depth (1–5 scale): whether caption provides specific critique rather than generic description; and (3) Overall Quality (1–5 scale): holistic assessment of caption usefulness. Inter-annotator agreement measured by Fleiss' kappa yielded $\kappa = 0.76$, indicating substantial agreement.

Results in Table 12 show ACSFM (BLIP+ACSFM and ViT-GPT2+ACSFM variants) significantly outperforms base models and LLaVA-1.5 across all dimensions. Specifically, BLIP+ACSFM achieves highest average scores (Aesthetic Relevance: 4.2, Evaluative Depth: 4.1, Overall: 4.3), substantially exceeding BLIP baseline (3.1, 2.8, 3.0) and LLaVA-1.5 (2.9, 2.6, 2.8). This confirms incorporating aesthetic sentiment feedback meaningfully improves caption quality from human perspective, particularly generating attribute-specific evaluative commentary rather than generic descriptions. LLaVA-1.5's lower performance, despite scale, reinforces that task-specific architectural design is critical for aesthetic evaluation beyond general-purpose vision-language capabilities.

**Table 12.** Human evaluation results comparing aesthetic caption quality across models. Three expert evaluators rated generated captions on a 5-point Likert scale across three dimensions: Aesthetic Relevance, Evaluative Depth, and Overall Quality. Inter-annotator agreement measured by Fleiss' kappa yielded $\kappa = 0.76$, indicating substantial agreement. Results are averaged over 100 test samples with standard deviations reported in parentheses. Bold values indicate the best-performing result for each evaluation dimension.

| Model | Aesthetic Relevance | Evaluative Depth | Overall Quality |
|---|---|---|---|
| BLIP | 3.1 (0.8) | 2.8 (0.7) | 3.0 (0.7) |
| **BLIP+ACSFM (ours)** | **4.2 (0.6)** | **4.1 (0.5)** | **4.3 (0.5)** |
| ViT-GPT2 | 3.3 (0.7) | 3.0 (0.8) | 3.2 (0.7) |
| **ViT-GPT2+ACSFM (ours)** | **4.0 (0.6)** | **3.9 (0.6)** | **4.1 (0.5)** |
| LLaVA-1.5 | 2.9 (0.9) | 2.6 (0.8) | 2.8 (0.8) |

### *5.8. Failure Case Analysis*

While models demonstrate strong overall performance, qualitative analysis of 50 systematically sampled errors reveals three primary failure modes: (1) Ambiguous Aesthetic Judgments —models struggle capturing nuanced sentiment balance when ground truth contains contradictory evaluations (e.g., nice composition but poor execution as shown in Figure 9), often oversimplifying to uniformly positive or negative sentiment; (2) Rare Aesthetic Attributes —infrequently discussed attributes (e.g., bokeh quality) receive neutral sentiment or fail detection entirely, with attributes appearing in <1% of training comments showing 23% lower F1 scores, indicating need for targeted data augmentation or few-shot learning; (3) Textual Noise Sensitivity —photography jargon, sarcasm, and informal expressions (e.g., blown out highlights misclassified as positive) confuse sentiment classification despite data cleaning. These failure modes collectively account for approximately 15% of ACSA errors and 18% of IAA prediction variance, suggesting targeted improvements in handling sentiment ambiguity, rare attributes, and domain-specific linguistic nuance could yield meaningful performance gains.

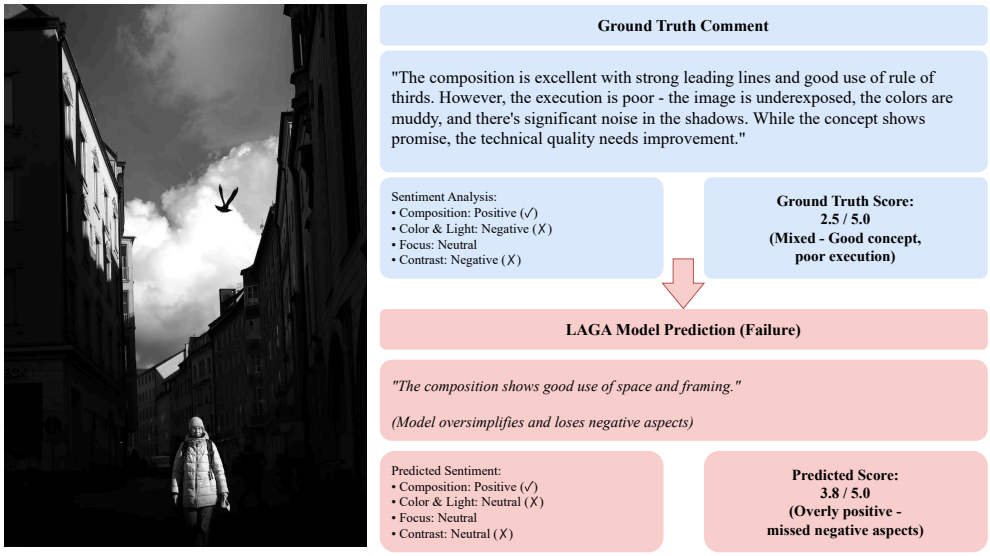


**Figure 9.** Example failure case demonstrating ambiguous aesthetic judgment with mixed sentiment that LAGA struggles to capture accurately. Model tends to oversimplify contradictory evaluations into uniform sentiment predictions.

## 6. Case Study

Figure 10 illustrates predicted scores for randomly selected 8 images using various models. For clarity, we separately selected one text-based model, one visual model, one multimodal-enhanced model, alongside proposed LAGA, comparing performance against ground truth scores. Figure 11 provides comparative analysis of aesthetic caption generation using multimodal models. Notably, both proposed LAGA and BLIP(+ACSFM) closely align with ground truth scores, demonstrating high accuracy. Additionally, ACSFM-generated comments exhibit strong similarity to ground truth evaluations, further validating effectiveness.

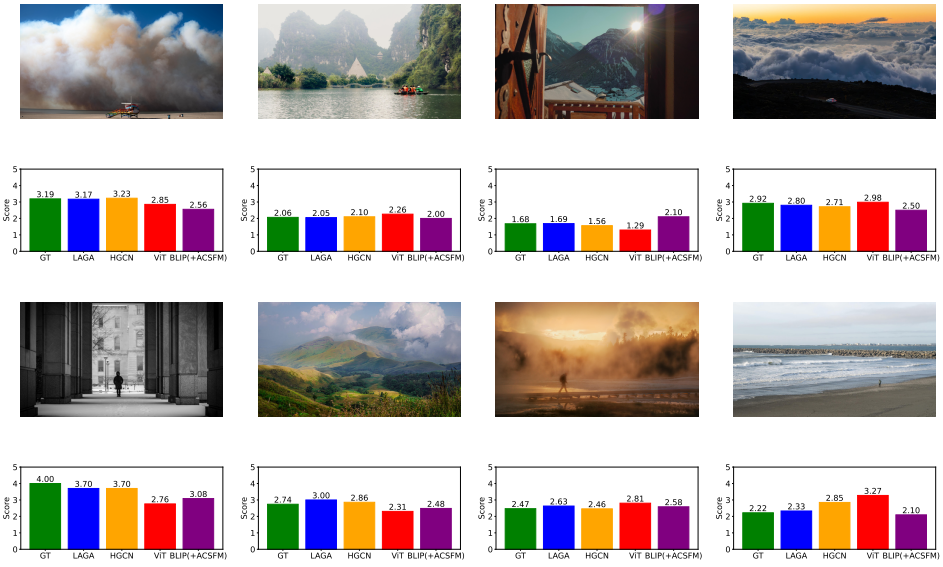


**Figure 10.** Case study comparing aesthetic scoring predictions across models. GT represents ground truth aesthetic scores. Graph demonstrates LAGA and BLIP(+ACSFM) achieve predictions closely aligned with ground truth values.



*GroundTruth: Composition is strong, with path leading into frame and trees occupying foreground, midground and background. Backlighting creates beautiful image, two colors—orange/brown and light green—add to aesthetics. I like it considerably.*

**BLIP(+ACSFM):** Trees with no real subject is great. I'd love seeing it from different perspective. It's close, but not particularly interesting. Two subjects are lost in frame, leading lines draw eye away from subject. Highlights are blown out and shadows are blown out. Highlights are blown, so bracketing and blending could help with exposure. Saturation is great, so keep shooting and editing.

**BLIP:** Nice arche de la defense. I've never seen it from this perspective, maybe bit too dark but that may just be what you were going for. It's great picture.

**ViT-GPT2(+ACSFM):** I like shot, but think it would be better if you could get closer to ground and made it bit more symmetrical. I also feel like there's bit too much going on in foreground but that's just me.

**ViT-GPT2:** I like photo, but think it would be better if there was more detail in sky and less of sky. Also, I think it's bit too dark for my taste.

**Figure 11.** Case study demonstrating comparative results for aesthetic caption generation across models. Generated captions from ACSFM variants exhibit stronger alignment with ground truth evaluations.

## 7. Conclusions

We introduce Textual Aesthetic Sentiment Labeling Pipeline (TASLP), an advanced methodology for annotating high-quality multimodal aesthetic data. Using TASLP, we construct Reddit Multimodal Sentiment Dataset (RMSD), expanding existing benchmarks by incorporating diverse aesthetic dimensions and fine-grained annotations capturing text-image interactions. RMSD addresses key gap in multimodal sentiment and aesthetic analysis, providing robust benchmark for tasks including image aesthetic assessment (IAA), multimodal sentiment analysis, and aesthetic-driven caption generation.

Building on RMSD, we propose Aesthetic Category Sentiment Analysis (ACSA) task for granular evaluation of aesthetic attributes through sentiment-aware perspective. To tackle ACSA, we introduce Layered Aesthetic Graph Convolutional Assessment Network (LAGA), a hierarchical model capturing complex aesthetic patterns. Experiments show LAGA consistently outperforms existing methods in IAA, setting new benchmark. Additionally, we develop Aesthetic Category Sentiment Fused Multimodal Model (ACSFM), integrating

multimodal feature fusion to enhance aesthetic sentiment classification and caption generation, highlighting benefits of sentiment-aware representations in computational aesthetics.

### 7.1. Limitations

Despite these advancements, our work has three important limitations: (1) Dataset Source Dependency—RMSD exclusively relies on r/photocritique, introducing biases in cultural perspectives (predominantly Western aesthetics), stylistic coverage (critique-focused sampling), and demographic representation (Reddit's user base), limiting generalizability to broader aesthetic contexts and non-Western traditions; (2) LLM Annotation Biases—reliance on GPT-4o for comment rewriting and sentiment labeling introduces systematic biases despite quality controls (Fleiss' kappa 0.78–0.83), potentially causing linguistic homogenization and subtle semantic drift from original intent; (3) Limited Aesthetic Domain Coverage—RMSD focuses on photographic aesthetics with four photography-centric attributes (Color & Light, Composition, Focus, Contrast) that may not generalize to other visual domains (paintings, illustrations, architectural design) requiring different aesthetic criteria.

### 7.2. Future Directions

To address these limitations and extend contributions, we propose three research directions: (1) Dataset Diversification—expand data collection beyond r/photocritique to multiple photography communities (Flickr, 500px, Instagram), artistic repositories (WikiArt, DeviantArt), and non-Western platforms to capture broader aesthetic perspectives and cultural diversity; (2) Hybrid Human-AI Annotation—develop collaborative frameworks where LLMs provide initial labels refined by human experts, employing active learning to prioritize uncertain samples and bias detection metrics to mitigate homogenization and preserve authentic expression; (3) Cross-Domain Aesthetic Modeling—extend ACSA to accommodate domain-specific attributes (e.g., brushwork for paintings, line weight for illustrations) with domain adaptation techniques for transfer across visual domains requiring different aesthetic criteria.

**Author Contributions:** Conceptualization, K.L. and M.P.; Methodology, K.L. and H.X.; Software, K.L.; Validation, K.L. and J.Z.; Resources, J.Z.; Data curation, H.X.; Writing—original draft, K.L.; Writing—review & editing, K.L., H.X., J.Z. and M.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by National Natural Science Foundation of China (NSFC) under Grant No. U23A20316 and Grant No. 62072346, and by Joint Laboratory on Credit Technology.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** RMSD dataset and source code for TASLP pipeline, LAGA and ACSFM models are available upon reasonable request. Dataset is distributed under Creative Commons Attribution 4.0 (CC BY 4.0) license, fully complying with Reddit's content policy and r/photocritique community guidelines. We provide pre-trained model checkpoints, evaluation scripts, and detailed instructions ensuring experimental reproducibility.

**Conflicts of Interest:** Authors declare no conflicts of interest. Funders had no role in study design; data collection, analyses, or interpretation; manuscript writing; or publication decision.

## Abbreviations

The following abbreviations are used:

| | |
|---|---|
| IAA | Image Aesthetic Assessment |
| IAC | Image Aesthetic Captioning |
| ACSA | Aesthetic Category Sentiment Analysis |
| TASLP | Textual Aesthetic Sentiment Labeling Pipeline |
| RMSD | Reddit Multimodal Sentiment Dataset |
| LAGA | Layered Aesthetic Graph Convolutional Assessment Network |
| ACSFM | Aesthetic Category Sentiment Fused Multimodal Model |
| MACE | Multi-Aesthetic Visual Channel Encoder |
| TACD | Transformer-based Aesthetic Caption Decoder |
| ACG | Aesthetic Category Group |
| GCN | Graph Convolutional Network |
| LLM | Large Language Model |
| PLCC | Pearson Linear Correlation Coefficient |
| SRCC | Spearman's Rank Correlation Coefficient |
| BLEU | Bilingual Evaluation Understudy |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |

## References

1. Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; Wang, L. Mining latent structures for multimedia recommendation. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 3872–3880.
2. May, A.; Chaintreau, A.; Korula, N.; Lattanzi, S. Filter & follow: How social media foster content curation. In Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems, Austin, TX, USA, 16–20 June 2014; pp. 43–55.
3. Yang, Y.; Xu, L.; Li, L.; Qie, N.; Li, Y.; Zhang, P.; Guo, Y. Personalized Image Aesthetics Assessment with Rich Attributes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19861–19869.
4. Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A large-scale database for aesthetic visual analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2408–2415.
5. Zhou, Y.; Lu, X.; Zhang, J.; Wang, J.Z. Joint Image and Text Representation for Aesthetics Analysis. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 262–266.
6. Chang, K.Y.; Lu, K.H.; Chen, C.S. Aesthetic Critiques Generation for Photos. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3534–3543.
7. Gonzalez-Naharro, L.; Flores, M.J.; Martínez-Gómez, J.; Puerta, J.M. Novel groundtruth transformations for the aesthetic assessment problem. *Inf. Process. Manag.* **2023**, *60*, 103368. [CrossRef]
8. Obrador, P.; Schmidt-Hackenberg, L.; Oliver, N. The role of image composition in image aesthetics. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3185–3188.
9. Liang, S.; Wu, D.; Zhang, C. Enhancing image sentiment analysis: A user-centered approach through user emotions and visual features. *Inf. Process. Manag.* **2024**, *61*, 103749. [CrossRef]
10. Liu, Y.; Wen, Z.; Li, L.; Jing, P.; Fan, D. MIGF-Net: Multimodal interaction-guided fusion network for image aesthetics assessment. *Pattern Recognit.* **2026**, *172*, 112401. [CrossRef]
11. Xu, X.; Huang, S.; Li, W.; Wang, F. Attribute-guided aesthetic assessment for artistic images based on multimodal hybrid network. *Expert Syst. Appl.* **2026**, *299*, 129954. [CrossRef]
12. Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; Lee, Y.J. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. *arXiv* **2024**, arXiv:2407.03341.
13. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24185–24198.
14. Kao, Y.; Wang, C.; Huang, K. Visual aesthetic quality assessment with a regression model. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Québec City, QC, Canada, 27–30 September 2015; pp. 1583–1587.
15. Marchesotti, L.; Murray, N.; Perronnin, F. Discovering beautiful attributes for aesthetic image analysis. *Int. J. Comput. Vis.* **2015**, *113*, 246–266. [CrossRef]

16. Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.T.; Wang, J.Z.; Li, J.; Luo, J. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.* **2011**, *28*, 94–115. [CrossRef]
17. Jin, X.; Wu, L.; Zhao, G.; Li, X.; Zhang, X.; Ge, S.; Zou, D.; Zhou, B.; Zhou, X. Aesthetic Attributes Assessment of Images. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 311–319.
18. Nieto, D.V.; Celona, L.; Labrador, C.F. Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment. In Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, LA, USA, 28 November–9 December 2022.
19. Lu, X.; Lin, Z.; Jin, H.; Yang, J.; Wang, J.Z. Rating image aesthetics using deep learning. *IEEE Trans. Multimed.* **2015**, *17*, 2021–2034. [CrossRef]
20. Kong, S.; Shen, X.; Lin, Z.; Mech, R.; Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In Proceedings of the ECCV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 662–679.
21. Zhang, B.; Niu, L.; Zhang, L. Image composition assessment with saliency-augmented multi-pattern pooling. In Proceedings of the BMVC, Virtual, 23–25 November 2021.
22. Zhu, T.; Li, L.; Chen, P.; Wu, J.; Yang, Y.; Li, Y. Emotion-aware hierarchical interaction network for multimodal image aesthetics assessment. *Pattern Recognit.* **2024**, *154*, 110584. [CrossRef]
23. Zhang, X.; Xiao, Y.; Peng, J.; Gao, X.; Hu, B. Confidence-based dynamic cross-modal memory network for image aesthetic assessment. *Pattern Recognit.* **2024**, *149*, 110227. [CrossRef]
24. Celona, L.; Leonardi, M.; Napoletano, P.; Rozza, A. Composition and Style Attributes Guided Image Aesthetic Assessment. *Trans. Img. Proc.* **2022**, *31*, 5009–5024. [CrossRef] [PubMed]
25. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
26. Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; Yang, F. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 10041–10051.
27. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
28. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 34892–34916.
29. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv* **2023**, arXiv:2308.12966.
30. Zhao, Y.; Yan, L.; Sun, W.; Xing, G.; Wang, S.; Meng, C.; Cheng, Z.; Ren, Z.; Yin, D. Improving the Robustness of Large Language Models via Consistency Alignment. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024), Torino, Italy, 20–25 May 2024; pp. 8931–8941.
31. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
32. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805. [CrossRef]
33. Team, Q. Qwen2 technical report. *arXiv* **2024**, arXiv:2412.15115.
34. OpenAI. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
35. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
36. Wang, Z.; Chen, J.; Chen, J.; Chen, H. Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics* **2024**, *129*, 7359–7384. [CrossRef]
37. Gokcimen, T.; Das, B. Topic modelling using bertopic for robust spam detection. In Proceedings of the 2024 12th International Symposium on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, 29–30 April 2024; IEEE: New York, NY, USA, 2024; pp. 1–5.
38. Ding, X.; Zhou, J.; Dou, L.; Chen, Q.; Wu, Y.; Chen, A.; He, L. Boosting Large Language Models with Continual Learning for Aspect-based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 4367–4377.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers), pp. 4171–4186.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.

41. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
42. Liu, J.; Teng, Z.; Cui, L.; Liu, H.; Zhang, Y. Solving Aspect Category Sentiment Analysis as a Text Generation Task. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominicana, 7–11 November 2021; pp. 4406–4416.
43. Cai, H.; Tu, Y.; Zhou, X.; Yu, J.; Xia, R. Aspect-Category based Sentiment Analysis with Hierarchical Graph Convolutional Network. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 833–843.
44. Liang, B.; Su, H.; Yin, R.; Gui, L.; Yang, M.; Zhao, Q.; Yu, X.; Xu, R. Beta Distribution Guided Aspect-aware Graph for Aspect Category Sentiment Analysis with Affective Knowledge. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominicana, 1–11 November 2021; pp. 208–218.
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
46. Tong, J.; Zhang, G.; Kong, P.; Rao, Y.; Wei, Z.; Cui, H.; Guan, Q. An interpretable approach for automatic aesthetic assessment of remote sensing images. *Front. Comput. Neurosci.* **2022**, *16*, 1077439. [CrossRef] [PubMed]
47. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5148–5157.
48. Bithel, S.; Bedathur, S. Evaluating Cross-Modal Generative Models Using Retrieval Task. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 1960–1965.
49. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
50. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318.
51. Lin, C.Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, 25–26 July 2004*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.