SAGA: SINGLE-SEQUENCE COUPLING OF DESCRIPTION AND SCORING VIA SCENE-ATTRIBUTE GRAPHS BASED AESTHETIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image quality assessment has evolved from a simple numerical regression to integrate score prediction with explanatory text generation. Existing methods usually treat these two objectives as completely independent tasks and train them in a decoupled manner, which may break the inherent consistency relations between the two tasks. Yet, directly modeling both tasks within a single generation sequence often leads to optimization conflicts and training instability due to the heterogeneity of target formats and imbalanced supervision signals. In this paper, we propose a novel **S**cene-**A**ttribute **G**raphs based **A**esthetic learning (SAGA) framework. Dual expert branches are fused within SAGA, based on their task-specific aesthetic experience latent spaces, to enhance the fusion model's capacity of capturing intricate aesthetic preferences. Thus, SAGA can effectively unify description generation and quality scoring within a single sequence. Extensive experiments show that SAGA significantly outperforms the state-of-the-art methods on both description generation and quality scoring, while exhibiting more balanced multitask performance. Furthermore, compared with models trained separately, SAGA maintains strong reasoning capability, while achieving notably lower inference latency.

1 INTRODUCTION

Image quality assessment (IQA) aims to automatically evaluate the perceptual quality of images in a manner consistent with human subjective judgments. It plays a fundamental role in a wide range of applications (Fang et al., 2020a; Wu et al., 2023), including image restoration, compression, generation, and content recommendation. As visual content continues to proliferate across diverse platforms, the demand for accurate, efficient, and interpretable quality assessment has become increasingly pressing. Beyond predicting perceptual quality scores, modern IQA systems are also expected to provide human-understandable explanations that reveal the underlying rationale of quality scoring, especially the causes of quality degradation. Such interpretability is crucial for enhancing user understanding, facilitating downstream decision-making, and enabling more effective optimization in practical applications.

Early IQA methods rely on handcrafted statistical features and classical metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) (Wang et al., 2004) to quantify perceptual degradation. Recently, deep-learning-based approaches build quality-aware representations directly from data, achieving notable performance gains over traditional methods. However, these methods are inherently limited to producing a single numerical score, offering no insight into why an image receives it. This lack of interpretability restricts their applicability in scenarios where diagnostic feedback or quality explanation is required. Moreover, these methods also face significant challenges in out-of-distribution (OOD) generalization.

The emergence of vision-language models (VLMs) has opened up new directions for IQA. Recent studies (You et al., 2025; Wu et al., 2023) leverage their strong visual understanding capabilities to directly predict image quality scores, achieving superior performance across multiple benchmarks (Fang et al., 2020b; Hosu et al., 2020; Lin et al., 2019) with the help of the rich pre-trained representations. In addition, other works (Cao et al., 2025; Li et al., 2026; Wu et al., 2024b; 2026) further explore the text generation ability of VLMs, and build IQA systems to perform both quality scoring

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

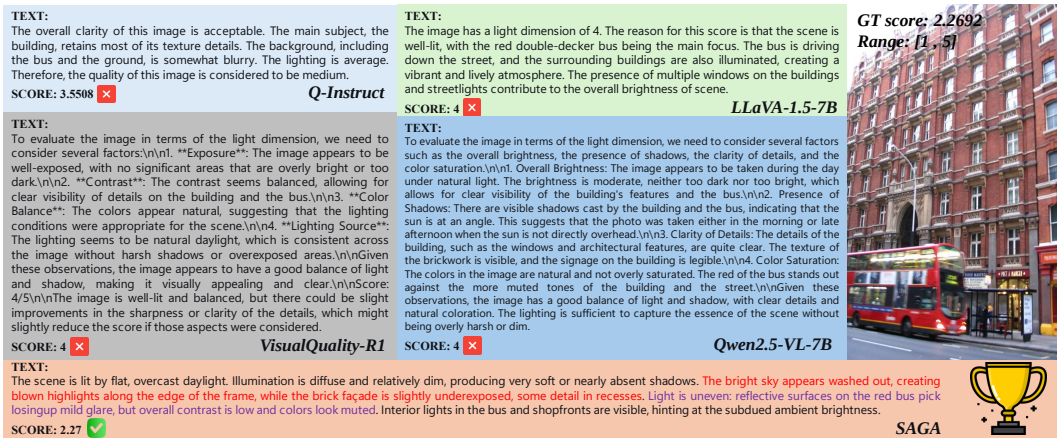

Figure 1: SAGA outperforms competing methods in both quality scoring and description generation tasks. The complete text/score outputs of evaluated models are presented for the image from PARA.

and description generation for the sake of model interpretability and human understandability. However, most of these methods treat scoring and description as entirely independent tasks, training and inferring them in a decoupled manner with separate task-specific prompts. This paradigm inherently neglects the intrinsic semantic dependence between the two tasks: an image described with a series of quality issues in the text output tends to receive a low score, and conversely, a low score implies a notable description of the quality degradation in the image. In essence, the description manifests the reasoning basis for the score, while the score serves as the quantitative summary of the description.

Motivated by the above considerations, we propose a coupled modeling approach, Scene-Attribute Graphs based Aesthetic learning (SAGA), to integrate description generation and quality scoring into a single generation sequence under a unified instruction, thus enabling a more consistent and comprehensive multitask IQA system. To resolve the joint modeling conflicts arising from heterogeneous task formats and imbalanced supervision signals, SAGA introduces dual expert branches within a multimodal foundation model, each dedicated to one task. Built upon shared visual-textual representations, the two branches learn task-specific transformations while being jointly optimized, facilitating knowledge sharing and strengthening the intrinsic correlation between the two tasks. Furthermore, to enhance the model's ability to perceive images, each expert branch is built upon an aesthetic experience latent space, where diverse image scenes and eval attributes are explicitly modeled as learnable nodes in a bipartite graph. A Graph Neural Network (GNN) (Scarselli et al., 2008) aggregates scene-attribute associations to produce aesthetic empirical features, providing the foundation model with rich aesthetic priors for fine-grained image assessment. Extensive experiments on SPAQ (Fang et al., 2020b), AIGCIQA2023 (Wang et al., 2023), and PARA (Yang et al., 2022a) demonstrate that SAGA achieves state-of-the-art accuracy on both tasks, while maintaining more balanced multitask performance. Figure 1 illustrates the outstanding performance of SAGA in both description generation and quality scoring on an image from PARA. Moreover, compared with separately trained SPAQ, AIGCIQA2023, and PARA models, SAGA preserves strong reasoning capability with lower inference latency, reducing per-sample inference time by 8.9%, 5.9%, and 7.2%, respectively. In summary, the main contributions of this work are as follows:

- We propose SAGA, a novel joint learning framework that unifies description generation and quality scoring into a single sequence, effectively addressing the joint optimization conflicts between these two heterogeneous tasks while enhancing cross-task consistency.
- We present aesthetic experience latent spaces based on scene-attribute GNN modeling, which capture intricate aesthetic associations between scenes and attributes, and provide rich aesthetic priors for fine-grained IQA.
- Extensive experiments show that SAGA achieves state-of-the-art accuracy on both tasks with more balanced multitask performance, and lower inference latency compared to separately trained models.

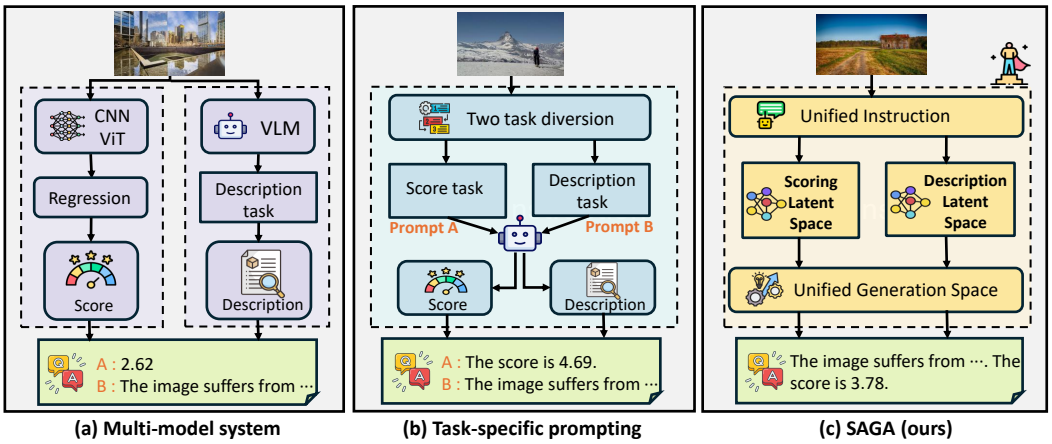

Figure 2: Comparison of different methods, including the multi-model system, task-specific prompting, and SAGA, which jointly models the two heterogeneous tasks in a unified sequence.

2 Related Work

2.1 Score-based Methods

Score-based methods have evolved through several stages. Early approaches rely on handcrafted statistical features and traditional metrics such as SSIM (Wang et al., 2004) to quantify perceptual quality (Ma et al., 2017; Mittal et al., 2012a;b). Subsequently, deep-learning-based methods (Liu et al., 2017; Talebi & Milanfar, 2018; Ke et al., 2021; Sun et al., 2022) learn quality-aware representations directly from data, achieving significant performance gains. More recently, VLM-based approaches such as Q-Align (Wu et al., 2023)] and DeQA-Score (You et al., 2025) harness the visual understanding capabilities of VLMs for quality prediction. However, these methods exclusively focus on score regression, while overlooking the generative potential of VLMs for quality description.

2.2 Description-based Methods

Recent efforts leverage VLMs to provide interpretable quality descriptions beyond numerical scores. Q-Bench (Wu et al., 2024a) and Q-Instruct (Wu et al., 2024b) improve models' low-level perceptual capabilities through carefully constructed visual instruction data. Co-Instruct (Wu et al., 2024c) enables a model to provide open-ended quality comparison conclusions and explanations in multi-image comparison scenarios. Q-Insight (Li et al., 2026) and VisualQuality-R1 (Wu et al., 2026) further incorporate reinforcement learning techniques. Q-Insight designs reward functions based on GRPO (Shao et al., 2024) to optimize both scoring and perception tasks separately. VisualQuality-R1 constructs ranking rewards by performing multiple quality estimations on paired images. In addition, Refine-IQA (Jia et al., 2026) employs a two-stage GRPO strategy with multi-task rewards to enhance the model's perception of low-level features. However, these methods treat description and scoring as independent tasks through separate modeling (one-stage/two-stage). How to jointly model these heterogeneous tasks while promoting cross-task consistency within a unified sequence remains an open challenge.

3 Methodology

3.1 Problem Formulation

We consider two heterogeneous tasks: description generation and quality scoring. Given an image, the description task generates a free-form text, while the scoring task predicts a score; the former is generative, requiring fine-grained perception and fluent generation, while the latter is discriminative, requiring compact and robust judgment. As shown in Figure 2, unlike existing approaches that handle the two tasks separately, SAGA jointly models both in a unified sequence, where task-specific experts

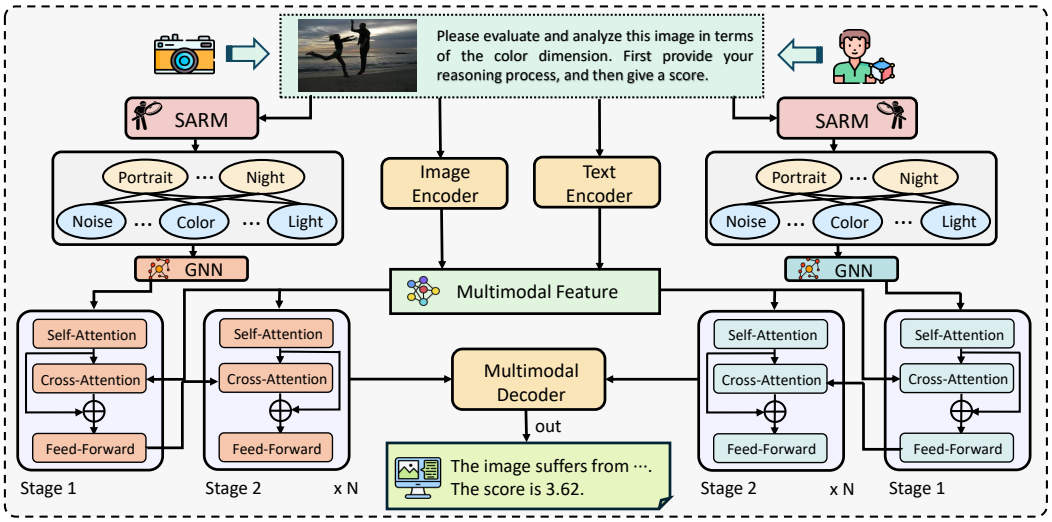

Figure 3: Overview of SAGA. Modules with the same color are shared between the two experts; in particular, the multimodal decoder is jointly updated by both tasks.

learn their respective capabilities and shared modules enable cross-task knowledge transfer, yielding a more consistent, balanced, and compact IQA system.

3.2 SAGA

We first provide an overview of proposed framework, SAGA. As shown in Figure 3, we construct two expert branches within the visual-textual foundation model, which are respectively designed for auxiliary description generation (left) and quality scoring (right). Specifically, an image-text input is first processed by encoders to extract features and construct multimodal features. Within each branch, the Scene Attribute Recognition Module (SARM) performs scene recognition based on the image and extracts attributes according to the query. Based on the identified (scene, attribute) pair, a relevant subgraph is extracted from the pre-constructed bipartite graph, and processed by a GNN to yield aesthetic empirical features. The multimodal features are then fed into the two branches, where they are dynamically fused with the aesthetic empirical features, and the resulting representations are then used for subsequent inferences.

Bipartite Graph Construction. Inspired by works (You et al., 2020; Chen & Zhang, 2020; Zhang et al., 2022) that use graphs to model cross-data relationships, we construct a bipartite graph $G = (V, E)$ between scenes and attributes to model their aesthetic correlations. The node set $V = V_S \cup V_A$ consists of n_S scene nodes from set $V_S = \{s_1, s_2, \ldots, s_{n_S}\}$ and n_A attribute nodes from set $V_A = \{a_1, a_2, \ldots, a_{n_A}\}$ with $n_S > 0$, $n_A > 0$ and $V_S \cap V_A = \emptyset$. The edge set $E = \{e_{s,a} \mid s \in V_S, a \in V_A\}$. Each edge $e_{s,a}$ represents a connection between scene s and attribute a. In this work, without loss of generality, we define V_S that consists of ten scenes: Natural Landscape, Urban Street Scene, Architecture/Landmarks, Transportation Scene, Indoor/Home Environment, Office/Study Environment, Commercial/Public Spaces, People/Portraits, Animals/Plants, and Still Life/Products/Food. V_A includes nodes for attributes defined in the evaluation data.

Scene Attribute Recognition Module. SARM is used to perform scene and attribute recognition on the input image-text data and output structured (scene, attribute) pairs. Specifically, scene recognition is conducted using a finetuned ViT (Dosovitskiy et al., 2021), whose output features are utilized for a 10-class classification task. Attribute extraction is performed through keyword matching with the text inputs. This module is implemented in a plug-and-play manner, decoupled from VLMs, and trained independently.

Aesthetic Experience Latent Space. Image quality/aesthetic assessments are inherently scene- and attribute-dependent: for instance, the same level of noise may yield different perceptual acceptability across night and portrait scenes. Such nuanced judgments reflect accumulated aesthetic experience

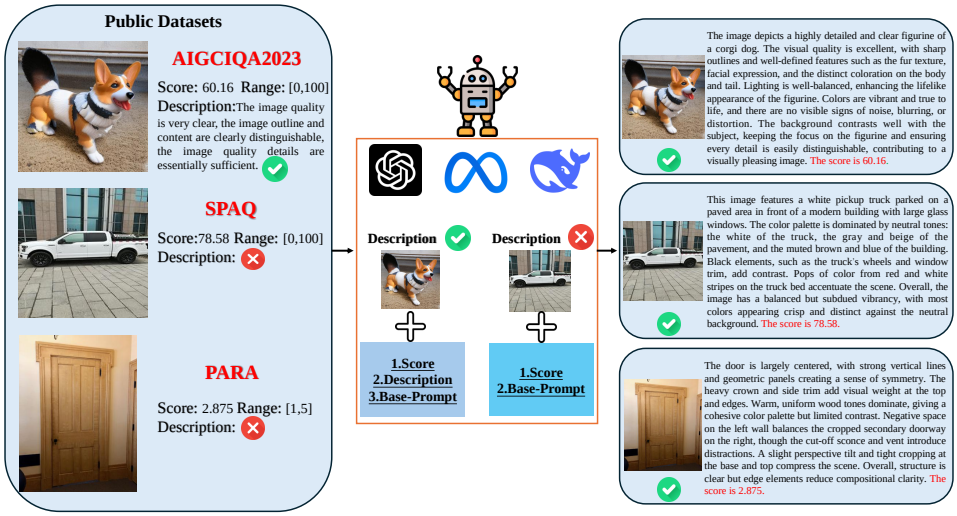

Figure 4: Data construction. Gemini 3.5 Flash is used in our experiments to construct evaluation data.

developed through extensive human evaluations. To structurally model the human aesthetic judgment process, we introduce an aesthetic experience latent space based on scenes and attributes. Given the graph G, SAGA uses a multi-layer GNN to capture aesthetic correlations between scenes and attributes, and output aesthetic empirical features K. It simply estimates K as below:

$$K = \text{GNN}(eid_{s,a}, eattr_{s,a}, nids_{s,a})$$

where $eid_{s,a}$ and $eattr_{s,a}$ denote the index and attribute of edge $e_{s,a}$, and $nids_{s,a}$ denotes the indices of nodes s and a. Specifically, within the GNN, we define a learnable embedding layer as the feature matrix, where each row represents the embedding vector of a node. We then use a 4-layer Graph Convolutional Network (GCN) (Kipf & Welling, 2017) to propagate update messages. In each batch, (scene, attribute) pairs identified by SARM are converted to their corresponding indices. These indices are then used to extract the corresponding subgraphs from G, which are fed as the inputs to GNN. After outputting the updated matrix, SAGA extracts the features of corresponding nodes based on the indices and concatenates them to construct the final aesthetic empirical features K.

To efficiently inject aesthetic experiences, we adopt an external modular fusion, where the multimodal features are fused with empirical features prior to being fed into the decoder. Specifically, we extract the multimodal features from the foundation model and fuse them with the empirical features K using attention mechanisms. This fusion process consists of two stages. The first stage consists of one self-attention layer and one cross-attention layer, which aligns the features. In this stage, the empirical features serve as the query, and the multimodal features serve as the key and value. The second stage consists of multiple layers of self-attention and cross-attention, where the query consists of the multimodal features, and the key and value are the feature output from the first stage. Ultimately, the final feature output from this fusion module is used for subsequent model propagation.

3.3 Optimization

In SAGA, the multimodal features are passed through two expert branches to generate description features and scoring features for downstream reasoning, respectively. During optimization, we adopt a joint loss for training; the aesthetic experience latent spaces corresponding to different tasks are updated independently, while the backbone network is shared across tasks and jointly optimized, facilitating knowledge sharing and strengthening the intrinsic correlation between the two tasks.

Description Generation Loss. We adopt the standard next-token prediction objective for description generation. Given the target description sequence $\mathbf{y}^d = (y_1^d, y_2^d, \ldots, y_T^d)$ with $T > 1$, the generation

loss is formulated as:

$$\mathcal{L}_{desc} = -\sum_{t=1}^{T} \log P(y_t^d \mid y_{<t}^d, \mathbf{x})$$

where $\mathbf{x}$ denotes the input image-text pair.

Quality Scoring Loss. Similarly, we formulate quality scoring as a next-token prediction task, where the target is the score sequence $\mathbf{y}^s = (y_1^s, y_2^s, \ldots, y_M^s)$ with $M > 1$:

$$\mathcal{L}_{score} = -\sum_{t=1}^{M} \log P(y_t^s \mid y_{<t}^s, \mathbf{x})$$

Then, the overall training objective is:

$$\mathcal{L} = \mathcal{L}_{desc} + \mathcal{L}_{score}$$

3.4 Data Construction

Existing datasets (He et al., 2022; Yang et al., 2022b; Fang et al., 2020b) often provide scores for corresponding attributes, but lack fine-grained descriptions, or contain almost no textual annotations. Therefore, to validate the effectiveness of SAGA, we follow ArtiMuse(Cao et al., 2025) and treat score labels as priors, leveraging VLMs to generate corresponding descriptions. As illustrated in Figure 4, we directly adopt the original annotated scores from datasets. For the datasets that only provide quality scores, we input the scores together with a basic prompt to guide the model in generating detailed image quality descriptions. For the datasets that contain both quality scores and limited textual annotations, we further incorporate the existing descriptions along with the scores and basic prompts to guide the model toward producing more fine-grained quality descriptions. Detailed prompt templates are provided in Appendix A. The datasets expanded with the generated descriptions are then used in the following evaluation experiments. The description length is uniformly set to 100 words for each dataset.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. To validate the generalizability of SAGA, we evaluate on three datasets spanning both the IAA and IQA domains: PARA, SPAQ, and AIGCIQA2023. PARA contains 7 attributes, SPAQ contains 5 attributes, and AIGCIQA2023 contains 2 attributes. Each dataset provides expert-annotated scores for the corresponding attributes. The score range of PARA is [1,5], while SPAQ and AIGCIQA2023 adopt a score range of [0,100]. Following the pipeline in Section 3.3, we construct label sequences containing fine-grained quality descriptions and accurate scores. For the scoring task, we adopt Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) as evaluation metrics. For the description task, we employ a weighted similarity metric that combines semantic information and character-level morphological information, with scores bounded in [0,2]. Specifically, the semantic component is computed by generating embeddings using BGE (Chen et al., 2024) and measuring the cosine similarity between the generated and ground-truth embeddings, while the character-level component is calculated based on the n-gram method (Papineni et al., 2002).

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct (Team, 2025) as the base model in SAGA. For an aesthetic experience latent space, the node feature dimension of a GNN is fixed at 300, and the number of attention layers in stage 2 is set to 4. For all components, we employ AdamW (Loshchilov & Hutter, 2019) as the optimizer with a learning rate of 1e-5. We maintain the same experimental settings across all experiments, which are conducted on 4 NVIDIA H800 80GB GPUs. For SARM training, we utilize images from KonIQ-10K (Hosu et al., 2020), SPAQ, PARA, and AIGCIQA2023 to train the ViT for scene recognition. Specifically, since the original datasets do not provide scene labels, we first employ Gemini 3 Flash to annotate the images with 10 scene categories, followed by manual spot-checking to verify the reliability of the annotations. Based on the generated scene labels, we then train the ViT with a learning rate of 1e-5 for 10 epochs. The final model achieves an accuracy of 97.87% on the test set. SARM is trained independently of the main system and integrated into the framework in a plug-and-play manner.

Table 1: Performance comparison of SAGA and competing methods on score regression tasks. The best and second-best results are highlighted in red and blue, respectively.

Model	PARA		SPAQ		AIGCIQA2023	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Models for General Purposes						
LLaVA-1.5-7B	0.217	0.256	0.329	0.228	0.149	0.151
Qwen2.5-VL-7B	0.726	0.733	0.475	0.454	0.052	0.532
Qwen3.8-Max	0.848	0.849	0.284	0.246	0.403	0.330
Gemini 3 Flash	0.827	0.817	0.562	0.550	0.349	0.290
GPT-5	0.806	0.814	0.510	0.469	0.548	0.518
GPT-5.4	0.736	0.729	0.305	0.294	0.539	0.519
Models for Specific Tasks						
Q-Instruct	0.813	0.799	0.768	0.750	0.631	0.650
VisualQuality-R1	0.713	0.720	0.541	0.510	0.399	0.525
Models for Joint Training						
Qwen2.5-VL-7B (SFT)	0.924	0.942	0.896	0.891	0.759	0.810
Qwen2.5-VL-7B (Q-Inight)	0.681	0.684	0.898	0.843	0.061	0.532
LLaVA-1.5-7B (SFT)	0.938	0.939	0.880	0.871	0.761	0.787
LLaVA-1.5-7B (Q-Inight)	0.239	0.266	0.298	0.225	0.163	0.158
SAGA (Ours)	**0.955**	**0.954**	**0.912**	**0.906**	**0.810**	**0.811**

4.2 Effectiveness of SAGA

Score Prediction. We evaluate a total of 13 models across the three datasets. These include general-purpose models (Qwen2.5-VL-7B, LLaVA-1.5-7B (Liu et al., 2024), Qwen3.8-Max, Gemini 3 Flash, GPT-5 and GPT-5.4), which are evaluated either via API or local calls. For models with undisclosed training details or trained with a disentangled objective (i.e., Q-Instruct and VisualQuality-R1), we evaluate their officially released checkpoints. The remaining models are trained on each dataset, including supervised fine-tuning (SFT), Q-Insight, and our proposed SAGA method. The results are reported in Table 1. Existing general-purpose models and methods designed for disentangled modeling generally perform poorly when simultaneously delivering quality descriptions and scores within single sequences. For instance, on SPAQ, the PLCC results range from as low as 0.284 for Qwen3.8-Max to only 0.768 for Q-Instruct. In contrast, SAGA demonstrates consistently superior performance, outperforming all the baseline methods across each benchmark. Particularly, on AIGCIQA2023, SAGA achieves an improvement of up to 0.749 in PLCC over the other methods and further exceeds the second-best model by 0.049.

Description Generation. We report the description generation performance in Table 2 for all attributes across different datasets. Herein, these models are evaluated using the weighted similarity metric proposed in Section 4.1. Experimental results demonstrate that SAGA not only exhibits excellent score prediction performance, but is also capable of generating accurate and detailed descriptions. SAGA achieves the best overall performance and outperforms competing methods on the vast majority of individual attributes.

User Study. We further conduct a human evaluation to validate the quality of the generated descriptions. Specifically, we randomly sample 100 examples from three datasets and recruit five users to independently evaluate the descriptions produced by different models along three dimensions: correctness, usefulness, and consistency. As shown in Table 3, SAGA achieves a preference of 69% on average across all users, which demonstrates that SAGA generates more accurate, useful, and consistent descriptions than existing baselines.

Table 3: Human eval on descriptions (preference %).

Evaluator	Qwen	SFT	Q-Insight	SAGA
User 1	5%	17%	6%	**72%**
User 2	4%	22%	9%	**65%**
User 3	6%	15%	8%	**71%**
User 4	9%	19%	4%	**68%**
User 5	3%	18%	10%	**69%**
Average	5.4%	18.2%	7.4%	**69%**

Table 2: Performance comparison of SAGA and competing methods on description generation tasks. The best and second-best results are highlighted in red and blue, respectively. The overall metrics are indicated by underlining.

Attribute	Qwen	Qwen (SFT)	Qwen (Q-Inight)	LLaVA	LLaVA (SFT)	LLaVA (Q-Inight)	Q-Instruct	VisualQuality-R1	SAGA
PARA									
Aesthetic	1.353	1.510	1.341	1.290	1.508	1.289	1.124	1.350	1.550
Composition	1.371	1.534	1.360	1.344	1.550	1.339	1.162	1.367	1.504
Quality	1.382	1.557	1.378	1.259	1.537	1.253	1.233	1.376	1.552
Color	1.194	1.488	1.189	1.178	1.488	1.172	1.018	1.189	1.443
Light	1.372	1.585	1.368	1.303	1.589	1.310	1.184	1.381	1.603
Content	1.259	1.493	1.255	1.306	1.498	1.308	1.051	1.247	1.513
Depth of Field	1.411	1.584	1.400	1.374	1.580	1.368	1.251	1.427	1.594
Overall	1.335	1.536	1.328	1.293	1.536	1.291	1.146	1.334	**1.537**
SPAQ									
Brightness	1.465	1.686	1.483	1.434	1.678	1.430	1.270	1.468	1.674
Contrast	1.527	1.688	1.471	1.429	1.681	1.428	1.287	1.531	1.711
Colorfulness	1.457	1.667	1.471	1.447	1.663	1.444	1.211	1.423	1.662
Sharpness	1.504	1.683	1.539	1.439	1.682	1.419	1.330	1.518	1.688
Noisiness	1.436	1.624	1.440	1.367	1.621	1.362	1.289	1.439	1.661
Overall	1.478	1.670	1.493	1.423	1.665	1.417	1.277	1.481	**1.679**
AIGCIQA2023									
Authenticity	1.461	1.615	1.461	1.405	1.613	1.405	1.244	1.465	1.617
Quality	1.503	1.633	1.503	1.395	1.627	1.395	1.353	1.496	1.632
Overall	1.482	1.624	1.482	1.400	1.620	1.400	1.298	1.480	**1.625**

4.3 Sensitivity Analysis and Ablation Studies

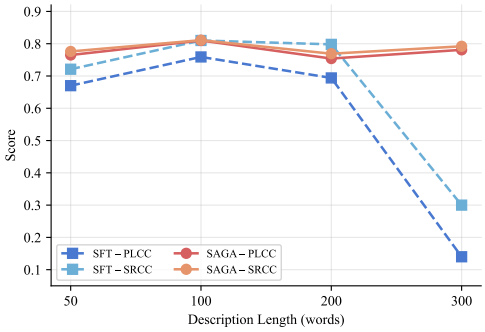

Figure 5: Effect of description length.

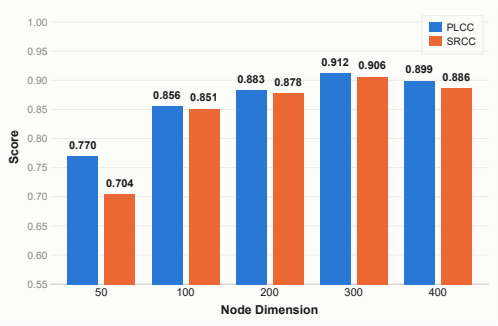

Figure 6: Effect of node dimension.

Datasets Variants. We construct four variants of AIGCIQA2023 by controlling the length of descriptions to 50, 100, 200, and 300 words, respectively. Experiments are conducted on Qwen2.5-VL-7B using the same settings, with the results shown in Figure 5. During joint modeling, variations in description length significantly affect a model's scoring ability, due to heterogeneous task objectives and imbalanced supervision signals. The baseline SFT method exhibits substantial performance fluctuations: when the description length increases to 300, its PLCC drops to as low as 0.140. This indicates that the baseline fails to resolve the optimization conflicts inherent in joint modeling. In contrast, SAGA maintains stable performance and consistently outperforms the baseline across different settings.

Node Dimension. To investigate the effect of node dimension on latent space modeling capability, we conduct experiments for SAGA on SPAQ under five settings, progressively increasing the node dimension from 50 to 400. As shown in Figure 6, within a certain range, higher dimensions lead to stronger modeling capacities of the aesthetic experience latent space and better optimization outcomes. Specifically, as the dimension grows from 50 to 300, PLCC improves from 0.770 to 0.912 and SRCC improves from 0.704 to 0.906. However, when the dimension reaches 400, both

Table 4: Architecture ablation studies. Metrics are reported as PLCC/SRCC/Description.

Setting	PARA	AIGCIQA2023	SPAQ
❄ w/o SARM	.889/.843/1.239	.737/.724/1.320	.799/.802/1.489
❄ w/o Graph	.918/.901/1.371	.748/.761/1.489	.886/.871/1.582
❄ Shared single branch	.801/.761/1.334	.614/.676/1.471	.742/.772/1.603
❄ Frozen decoder	.923/.919/1.412	.754/.767/1.508	.894/.888/1.610
🔥 w/o SARM	.886/.897/1.290	.768/.775/1.514	.755/.769/1.475
🔥 **Unfrozen decoder**	**.955/.954/1.537**	**.810/.811/1.625**	**.912/.906/1.679**

Table 5: Comparison between different training strategies. Best results are highlighted in red.

Dataset	Strategy	PLCC	SRCC	Text	Per Infer Time (s)	Total Time (s)
SPAQ	Separate	0.921	0.914	1.636	0.3149	944.7
	Joint (Ours)	0.912	0.906	1.679	0.2868	860.4
AIGCIQA2023	Separate	0.798	0.805	1.595	0.3535	339.36
	Joint (Ours)	0.810	0.811	1.625	0.3325	319.2
PARA	Separate	0.954	0.953	1.518	0.3756	1314.6
	Joint (Ours)	0.955	0.954	1.537	0.3484	1219.4

measurements decline, indicating that excessively high dimensions introduce feature redundancy that impairs generalization performance.

Architecture Design. We further conduct a comprehensive architecture ablation study to better quantify the contribution of each component. Specifically, we train a variant with the decoder frozen and perform module-wise ablation under this fixed setting. All variants share the same hyperparameter configuration to ensure a fair comparison. As shown in Table 4, we observe that (1) freezing the decoder prevents knowledge sharing through shared decoder gradients, thereby weakening the interaction between these two tasks; (2) removing SARM leads to a notable performance drop across all datasets, confirming that distinguishing and explicitly modeling scene attributes is essential ; (3) removing the Graph and relying solely on node embeddings results in degraded performance ;(4) merging the two branches into a shared single branch causes the most severe degradation.

Training Strategy. We compare SAGA with the separate training strategy on Qwen2.5-VL-7B, which models quality scoring and description generation as independent tasks. As shown in Table 5, in terms of scoring performance, SAGA achieves comparable results and even obtains superior PLCC/SRCC on AIGCIQA2023 and PARA. For description quality, SAGA consistently outperforms the separate training strategy, indicating that joint modeling effectively facilitates feature interaction and knowledge sharing. Furthermore, through single-sequence joint modeling, SAGA achieves the faster inference speeds, reducing per-sample time by 8.9% (SPAQ), 5.9% (AIGCIQA2023), and 7.2% (PARA), respectively.

5 Conclusion

In this paper, we proposed SAGA, a novel aesthetic-aware joint learning framework that unifies quality scoring and description generation within a single generation sequence. Based on the aesthetic experience latent spaces induced by scene-attribute knowledge graphs, SAGA effectively captures complex and subtle aesthetic preferences. Moreover, the joint optimization of the dual-expert branches alleviates the optimization conflicts caused by heterogeneous task formats and imbalanced supervision signals. In this way, SAGA sheds a new insight into joint optimization of heterogeneous objectives in IQA. Extensive experiments on SPAQ, AIGCIQA2023, and PARA demonstrate that SAGA achieves more balanced state-of-the-art performance on both scoring and description tasks. Moreover, compared with separately trained models, SAGA preserves strong reasoning capability with lower inference latency, validating the effectiveness and efficiency of joint modeling for IQA. As part of future work, we would like to extend this aesthetic learning framework for other vision-language tasks.

Ethics Statement

This work adheres to the ICLR Code of Ethics. No human subjects or animal experimentation was involved in this study. All datasets employed were collected and used in compliance with their respective terms of use, and no personally identifiable information was utilized at any stage of the research. We have carefully designed our experimental pipeline to avoid potential biases or discriminatory outcomes. The proposed method focuses solely on image quality assessment and description generation, posing no privacy or security concerns. We are committed to maintaining full transparency and research integrity throughout this work.

Reproducibility Statement

We are committed to ensuring the full reproducibility of our work. All training procedures, model architectures, hyperparameter settings, and hardware specifications are described in detail throughout the paper and Appendix. All datasets used in this study are publicly available, and the source code and model checkpoints will be released upon acceptance to facilitate replication and further research.

LLM Usage

Large Language Models (LLMs) were used solely to assist with language polishing and manuscript preparation, including sentence rephrasing, grammar correction, and improving readability. The LLM was not involved in any aspect of research ideation, methodology design, or experimental analysis. All scientific contributions, research concepts, and data analyses were entirely developed by the authors. The authors take full responsibility for all content in this manuscript.

References

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Jiayi Chen and Aidong Zhang. Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness. In *26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1295–1305, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3677–3686, 2020a.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3677–3686, 2020b.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, volume 6, pp. 22, 2022.

V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Zijian Chen, and Xiongkuo Min. Refine-iqa: Multi-stage reinforcement finetuning for perceptual image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22301–22309, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Jian Zhang, et al. Q-insight: Understanding image quality via visual reinforcement learning. *Advances in Neural Information Processing Systems*, 38:36802–36827, 2026.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 1040–1049, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 25: 2912–2925, 2022.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*, pp. 46–57. Springer, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, volume 2024, pp. 12547–12573, 2024a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25490–25500, 2024b.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377. Springer, 2024c.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Advances in Neural Information Processing Systems*, 38:88167–88190, 2026.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR), 00*, pp. 19829–19837, 2022a.

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

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33: 19075–19087, 2020.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14483–14494, 2025.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2418–2428, 2022.

A PROMPTS FOR DATA CONSTRUCTION

Table 6 presents the prompt templates used to generate fine-grained descriptions for each dataset. For PARA and SPAQ, the prompts incorporate the attribute name, its corresponding score, and the valid score range. For AIGCIQA2023, an additional brief description is included to provide supplementary context. All prompts instruct the VLMs to generate a detailed description of approximately 100 words conditioned on the expert rating and the given context.

Table 6: Prompt templates for data construction across different datasets.

Dataset	**Prompt Template**
PARA	This image has a <attribute> score of <attribute_score>, with a score range from 1 to 5. The higher the score, the better the performance in this dimension. Based on the expert rating and the image itself, please generate a detailed and accurate description in terms of <attribute>. The word count should be <num> words.
AIGCIQA2023	This image has a <attribute> score of <attribute_score>, with a score range from 0 to 100. The higher the score, the better the performance in this dimension. There is also a brief description: <des>. Based on the expert rating and the short description, please generate a detailed and accurate description in terms of <attribute>. The word count should be <num> words.
SPAQ	This image has a <attribute> score of <attribute_score>, with a score range from 0 to 100. The higher the score, the better the performance in this dimension. Based on the expert rating and the image itself, please generate a detailed and accurate description in terms of <attribute>. The word count should be <num> words.

B BALANCED MULTITASK PERFORMANCE.

We rank all models, according to their results in Table 1 and Table 2, from 1 (lowest) to 9 (highest). As illustrated in Figure 7, the horizontal and vertical axes represent the ranks for scoring and description ability, respectively. A balance line (y = x) indicates equal ranks on both tasks. Overall, the baselines exhibit varying degrees of task imbalance. Six models fall above the balance line, indicating stronger description ability than scoring ability, while two models fall below, suggesting the stronger scoring ability. In contrast, SAGA achieves the highest ranks on both tasks and lies on the balance line, which demonstrates SAGA effectively harmonizes both tasks during training, enhancing cross-task consistency.

C SARM

To evaluate the robustness of SARM under adverse conditions, we apply three types of perturbations to the input images: Gaussian blur, Gaussian noise, and uniform noise. As shown in Figure 8, SARM achieves 97.87% accuracy on the original test set and maintains strong performance under Gaussian blur (97.60%), demonstrating its resilience to spatial degradation. The accuracy slightly drops under Gaussian noise (94.94%) and uniform noise (93.84%), yet remains consistently above 93.8%. These results indicate that SARM exhibits reliable robustness across various perturbation scenarios.

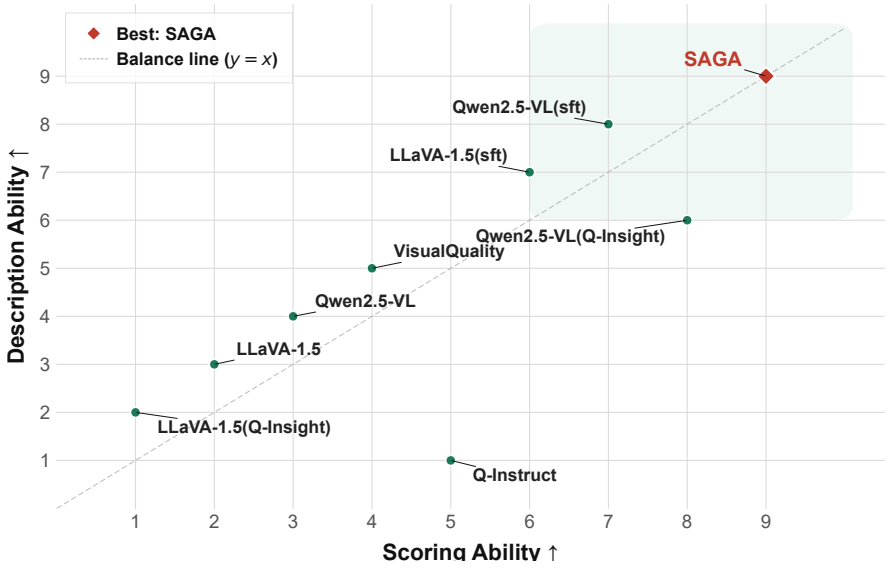

Figure 7: Balanced multitask performance of SAGA.

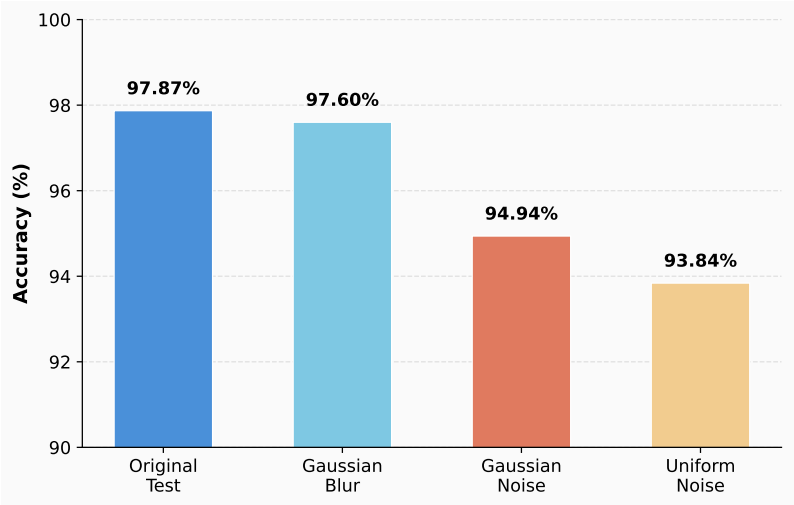

Figure 8: Scene recognition accuracy of SARM under different perturbation types.

D STATEMENT

In this paper, we use IQA as a general term encompassing both quality assessment and aesthetic evaluation. These two domains adopt similar formulations and model architectures, differ primarily in their annotated attributes. Our method is evaluated on datasets from both domains and achieves consistently strong performance, without being constrained by specific domain evaluation attributes. As a result, the aesthetic priors discussed in this paper cover a broader range of perceptual attributes, including but not limited to color, brightness, composition, and depth of field.

E LIMITATIONS

Due to data scarcity, we validated SAGA on three relevant and available datasets from distinct domains, which cover a limited set of scenes and attributes. However, the SAGA framework is data-agnostic and readily applicable to broader attributes and more diverse scenarios.