Advancing Comprehensive Aesthetic Insight with Multi-Scale Text-Guided Self-Supervised Learning

Yuti Liu*, Shice Liu*, Junyuan Gao, Pengtao Jiang, Hao Zhang, Jinwei Chen, Bo Li†

vivo Mobile Communication Co., Ltd, Shanghai, China
{kira66arik*, junyuangao577}@gmail.com; {liushice*, pt.jiang, haozhang, jinwei.chen, libra†}@vivo.com

Abstract

Image Aesthetic Assessment (IAA) is a vital and intricate task that entails analyzing and assessing an image's aesthetic values, and identifying its highlights and areas for improvement. Traditional methods of IAA often concentrate on a single aesthetic task and suffer from inadequate labeled datasets, thus impairing in-depth aesthetic comprehension. Despite efforts to overcome this challenge through the application of Multi-modal Large Language Models (MLLMs), such models remain underdeveloped for IAA purposes. To address this, we propose a comprehensive aesthetic MLLM capable of nuanced aesthetic insight. Central to our approach is an innovative multi-scale text-guided self-supervised learning technique. This technique features a multi-scale feature alignment module and capitalizes on a wealth of unlabeled data in a self-supervised manner to structurally and functionally enhance aesthetic ability. The empirical evidence indicates that accompanied with extensive instruct-tuning, our model sets new state-of-the-art benchmarks across multiple tasks, including aesthetic scoring, aesthetic commenting, and personalized image aesthetic assessment. Remarkably, it also demonstrates zero-shot learning capabilities in the emerging task of aesthetic suggesting. Furthermore, for personalized image aesthetic assessment, we harness the potential of in-context learning and showcase its inherent advantages.

1 Introduction

As artificial intelligence evolves, there's a growing demand for agents to mimic human perception and exhibit emotional responses to their surroundings. IAA emerges as a key area within this scope, and gauges images' aesthetic appeal akin to human judgment. Its complexity lies in its subjectivity, governed by factors like photographic subjects and personal experiences, which makes IAA a challenging endeavor.

In the last decade, IAA has been concretized into a variety of tasks. EAT (He et al. 2023b) predicts aesthetics based on a single human-assigned score per image—a task known as Aesthetic Scoring (AS). Meanwhile, CWS (Ghosal et al. 2019) assesses an image's aesthetic appeal directly through language, which is referred to Aesthetic Commenting (AC). Recently, Personalized Image Aesthetic Assessment (PIAA) has emerged as a burgeoning field, which aims to predict an individual's aesthetic preferences based on his historical image scoring. While effective in certain scenarios, approaches focusing on a single task often fail to address linkages between different tasks, suffering from overfitting to specific tasks. This realization has inspired us to prioritize holistic aesthetic analysis and comprehension in our research efforts.

Recently, MLLMs have demonstrated strong comprehension and reasoning abilities across various domains. Models such as VILA (Ke et al. 2023), Q-Align (Wu et al. 2023), and UNIAA (Zhou et al. 2024) have also attempted to utilize MLLMs for IAA to compensate for perceptual and reasoning processes. However, two major obstacles limit their effectiveness. First, these models rely solely on semantic features, neglecting a wealth of valuable aesthetic information. Second, despite efforts by Q-Align and UNIAA to construct aesthetic question-answer pairs for enhancement, the scarcity of labeled data and the presence of potentially mislabeled data continue to restrict their performance. Consequently, integrating comprehensive aesthetic information into MLLMs and developing a refined learning strategy to accurately leverage massive image data are essential.

In this paper, we propose Comprehensive Aesthetic Large language Model (CALM) which excels in various IAA tasks and demonstrates deep aesthetic comprehension and analytical skills in dialogues. Fig. 1 illustrates the functional differences between CALM and other IAA models.

Inspired by popular MLLMs, CALM incorporates a visual encoder, a Multi-scale Feature Alignment Module (MFAM) and a Large Language Model (LLM). Recognizing that mainstream visual encoders and LLMs excel at feature extraction and language expression, we have focused our efforts on the MFAM to ensure that the subsequent LLM can fully leverage a broader spectrum of aesthetic information provided by the visual encoder. To achieve this, we introduce a multi-scale text-guided self-supervised learning technique.

Specifically, the MFAM is designed to structurally access aesthetic features at multiple levels, while text-guided self-supervised learning enables the MFAM to benefit from unlabeled data. Unlike previous aesthetic self-supervised approaches that rely on score pseudo-labels, our method uses attribute-related textual pseudo-labels. This change ensures accurate learning and simplifies the integration of pseudo-

*These authors contributed equally.
†This author is the corresponding author.

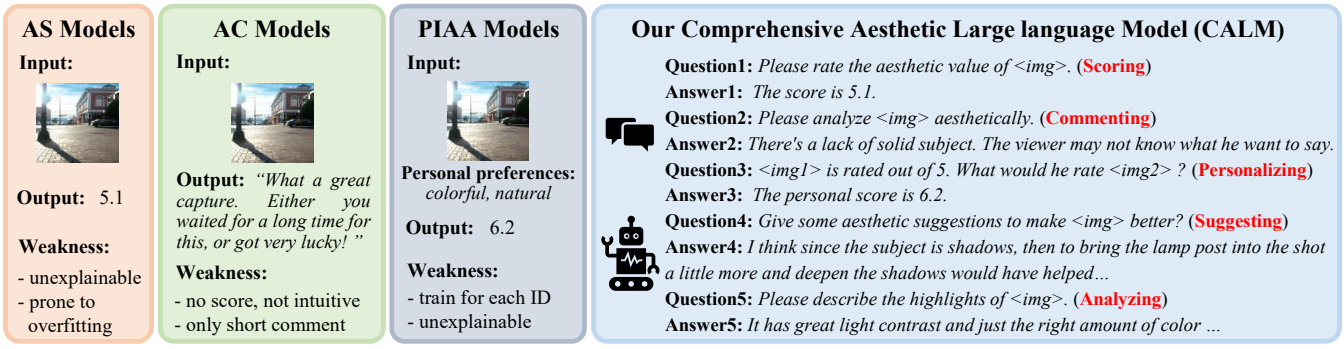

Figure 1: The functional comparison of our proposed CALM and other IAA methods.

labels when superimposing multiple augmentations on a single image. Additionally, we utilize a wider range of image augmentations, from low-level to high-level, to guarantee that more aesthetic elements are captured and learned.

To enhance holistic aesthetic insight, we developed various instruct-tuning techniques to adapt CALM to common aesthetic tasks, ultimately outperforming other approaches in AS, AC, and PIAA tasks. Moreover, CALM achieves comparable PIAA results through in-context learning at runtime, establishing a new paradigm for PIAA. Additionally, we are the first to define the aesthetic suggesting task, and CALM's zero-shot success in this task demonstrates its ability to grasp and comprehend aesthetic principles effectively.

The contributions of our work are concluded as follows:

♢ We propose CALM, a cutting-edge multi-modal large language model specialized in comprehending image aesthetics. Our extensive experiments demonstrate that CALM sets a new benchmark for AS, AC, and PIAA tasks.

♢ We have pioneered a multi-scale text-guided self-supervised learning technique that not only ensures multi-scale perception for MLLMs, but also effectively and efficiently leverages abundant unlabeled images for enhancement.

♢ The remarkable zero-shot capabilities of CALM are explored, particularly in in-context PIAA and providing aesthetic suggestions. These capabilities demonstrate CALM's comprehensive aesthetic insight and analytical prowess.

2 Related Work

Image Aesthetic Assessment involves algorithms that measure the visual appeal of images. Initially, convolutional neural networks (CNN) and transformers have been leveraged to refine aesthetic score predictions, such as TANet (He et al. 2022), ResNext (Hou et al. 2022b), DAT (Xia et al. 2022) and MaxViT (Tu et al. 2022). In order to regulate aesthetic features to refine scoring, Comm (Niu et al. 2022) and AesCLIP (Sheng et al. 2023) harness textual data and CLIP (Radford et al. 2021), respectively. Besides, language generation models for AC task have also emerged, such as Yeo (Yeo et al. 2021). Moreover, realizing the importance of personal tastes, models and the FLICKR-AES dataset (Ren et al. 2017) for PIAA are gaining traction. However, previous methods usually concentrate on a single aesthetic task so that they can barely really understand aesthetics.

Multi-modal Large Language Models achieve image content analysis by integrating visual features in LLMs. LLaVA-1.5 (Liu et al. 2023) and mPLUG-Owl2 (Ye et al. 2024) have showcased impressive image reasoning skills. In the realm of IAA, VILA employs CoCa (Yu et al. 2022) to explore zero-shot aesthetic judgement, while Q-ALIGN directly utilizes the original mPLUG-Owl2. UNIAA leverages ChatGPT to generate comments to fine-tune LLaVA-1.5. However, these methods do not modify the pre-existing MLLMs and rely on a limited number of constructed data, which may prevent a comprehensive aesthetic understanding. Consequently, it is vital to improve the structural design and functional learning for deeper aesthetic comprehension.

Multi-scale Aesthetic Perception is a key approach for promoting IAA. (Chen et al. 2020) combined multi-level spatial features and employed adaptive dilated CNNs, while Comm designed a module to process multi-scale features. EAT and ICAA (He et al. 2023a) incorporated interest points and delegate transformers, aligning better on specific scales. Drawing on these observations, we develop a technique for MLLMs that harnesses multi-scale features effectively.

Self-supervised Learning seeks to leverage large quantities of unlabeled data and artificially assigned pseudo-labels to enhance models' generalization (Chen and He 2021). In IAA, where expert annotation is often costly, self-supervised methods are particularly prevalent. (Sheng et al. 2020; Pfister et al. 2021) intuitively assigned lower aesthetic scores to augmented images for contrastive learning, generating score pseudo-labels. However, due to the still ambiguous factors influencing aesthetics and cases where depth-of-field blur can enhance aesthetic appeal, these methods risk producing inaccurate pseudo-scores. Moreover, these methods primarily focus on low-level data augmentations and require separate classifiers to regress scores, limiting their effectiveness.

3 Methodology

3.1 The Architecture of CALM

As represented in Fig. 2, CALM is composed of three principal elements: a visual encoder $g(\cdot)$ transforming an image $\mathrm{X_v}$ into a sequence of visual tokens $\mathrm{Z_v} = g(X_v)$; an MFAM $W(\cdot)$ converting visual tokens $\mathrm{Z_v}$ into vision-language tokens $\mathrm{H_v} = W(\mathrm{Z_v})$; an LLM $f(\cdot)$ that receives the vision-language tokens $\mathrm{H_v}$ and user instructions $\mathrm{X_q}$ to produce the

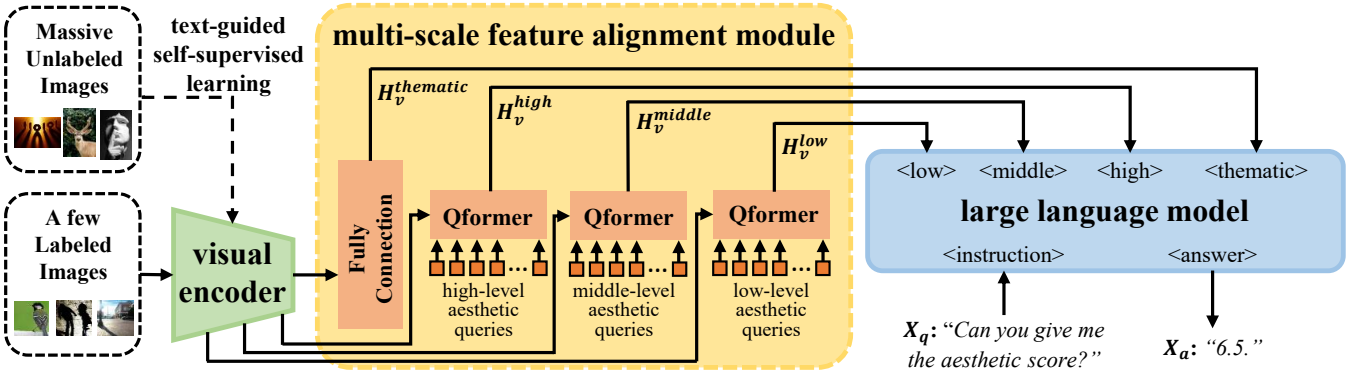

Figure 2: The proposed CALM includes a visual encoder, a multi-scale feature alignment module and a large language model.

relevant language responses $\mathrm{X_a} = f(\mathrm{H_v}, \mathrm{X_q})$.

Referring to most MLLMs, we employ the open-sourced ViT-L/14 as $g(\cdot)$ and Vicuna-7B (Chiang et al. 2023) as $f(\cdot)$ without any modification. For our proposed MFAM, we detail its structural design in Sec. 3.2 and its functional promotion via text-guided self-supervised learning in Sec. 3.3. Subsequently, we show how CALM simultaneously addresses various IAA tasks through two-stage instruct tuning in Sec. 3.4. What's more, only regression loss is employed to reduce the gap between $\mathrm{X_a}$ and the ground truth $\mathrm{X_{gt}}$.

3.2 Multi-scale Feature Alignment Module

(Jin et al. 2019) have revealed that image clarity and color schemes are encoded in lower-level features, while composition and impression requires higher-level features for interpretation. Although multi-scale features have been broadly explored in IAA, MLLMs, which typically process tokens from the last several layers of the visual encoder, lacks a structural basis for handling multi-scale features. Hence, we design the MFAM to emphasize multi-scale information.

We define four levels based on their positions in $g(\cdot)$, from shallow to deep sequentially named as low-, middle-, high-, and thematic-level. To preserve the original reasoning ability, we utilize fully connection to yield thematic-level features $\mathrm{H_v^{thematic}} \in \mathcal{R}^{N_v \times d_l}$, where N_v and d_l are the number of vision tokens and the dimension of language tokens, respectively. And then, three two-layer Qformers (Li et al. 2023) are introduced, which use cross attentions to make learnable queries pinpoint aesthetic features at the targeted levels. With $g(\cdot)$ offering 24 hidden state layers, we strategically tap into the 4th, 12th, and 24th layers to compute low-level features $\mathrm{H_v^{low}} \in \mathcal{R}^{N_{low} \times d_l}$, middle-level features $\mathrm{H_v^{middle}} \in \mathcal{R}^{N_{middle} \times d_l}$, and high-level features $\mathrm{H_v^{high}} \in \mathcal{R}^{N_{high} \times d_l}$, where N_{low}, N_{middle} and N_{high} denote the number of learnable queries at each level. The design of MFAM makes it effective and efficient to capture key aesthetic features, considering that the number of queries is much smaller than that of visual tokens.

3.3 Text-guided Self-supervised Learning

For the purpose of effectively unlocking the potential of abundant unlabeled image data to accurately enhance aesthetic perception, we propose text-guided self-supervised learning, which offers the following three advantages.

Firstly, we use accurate attribute pseudo-labels to replace flawed score pseudo-labels for self-supervision. Concretely, we introduce various image augmentation algorithms targeting attributes mentioned in (Jin et al. 2019), such as color and subject. During training, unlabeled images are randomly augmented in certain attributes and assigned the corresponding attribute pseudo-labels. For instance, if an image is blurred, its attribute pseudo-label is "the blurred image".

Secondly, we leverage a broader spectrum of data augmentations compared to previous aesthetic self-supervised methods. These are categorized into three types: degradation of image quality (e.g., Gaussian blur, impulse noise, JPEG compression, pixelate, motion blur, defocus blur), alteration of image color (e.g., brightness, saturation, contrast adjustments), and modification of image content (e.g., subject or non-subject object masking, foreground or background blurring). Subsequent experiments confirm that these image augmentations significantly enhance aesthetic insight.

Thirdly, we employ GPT-3.5 to generate various textual contrastive pseudo-labels, which eliminates the need for specialized classifiers in (Jin et al. 2019). Two examples are provided in the self-supervised pre-training part in Fig. 3. Additionally, multiple augmentations can be applied simultaneously with their textual pseudo-labels conveniently spliced into a cohesive target, increasing both the data volume and the variety of contrastive learning. For instance, if an image is blurred and added noise, the pseudo-label would be, "The first image is blurrier and noisier than the second".

3.4 Comprehensive Aesthetic Assessment

To achieve comprehensive aesthetic insight, we employ two-stage instruct tuning to adapt CALM to various aesthetic tasks, such as AS, AC, and PIAA. Specific instruction examples are shown in Fig. 3. The complete training cycle, illustrated in Fig. 4, encompasses pre-training and fine-tuning.

The pre-training stage consists of two parts that can be launched simultaneously. **Self-Supervised Pre-Training** encourages the three Qformers in the MFAM to learn aesthetic attributes in a self-supervised manner, utilizing unlabeled images from diverse sources, including AVA (Murray

Abbreviations for our used prompts: <**thematic**> for "*semantic-level features:* $H_v^{thematic}$", <**high**> for "*high-level features:* H_v^{high}", <**middle**> for "*middle-level features:* H_v^{middle}", <**low**> for "*low-level features:* H_v^{low}", <**image**> for their combination "<**thematic**>, <**high**>, <**middle**>, <**low**>".

Two examples of the instructions for the self-supervised pre-training:
Instruction 1: *"The first image is* <**high1**>, <**middle1**>*; the second image is* <**high2**>, <**middle2**>*."* **Answer 1:** *"The first image has a better composition compared to the second."*
Instruction 2: *"The first image is* <**low1**>, <**middle1**>*; the second image is* <**low2**>, <**middle2**>*."* **Answer 2:** *"The first image is blurrier than the second."*

An example of the instructions for the generic pre-training:
Instruction: *"The image is* <**thematic**>*."* **Answer:** *"A dog is running in the grass with his owner chasing him behind. The dog seemed very happy."*

An example of the instructions for the aesthetic commenting fine-tuning:
Instruction: *"The image is* <**image**>*. Please comment on this image aesthetically."* **Answer:** *"The focus adjustment was great, and the light was right. This is a very good photo."*

An example of the instructions for the aesthetic scoring and PIAA fine-tuning:
Instruction: *"The image is* <**image**>*. Please rate it aesthetically."* **Answer:** *"8.5."*

An example of the instructions for the in-context learning PIAA:
Instruction: *"Knowing that an image,* <**image1**>, is rated by 4*. What would he rate* <**image2**>*?"* **Answer:** *"4.5."*

An example of the instructions for the aesthetic suggestion task:
Instruction: *"What should be most improved for the image,* <**image**>*, to enhance its aesthetic value?"* **Answer:** *"It has a great composition, but the exposure could be increased..."*
For GPT-3.5 evaluation: *"Answer whether the following text is intended to convey increased brightness."*

Figure 3: Some instruction examples utilized throughout the entire training process and across various tasks.

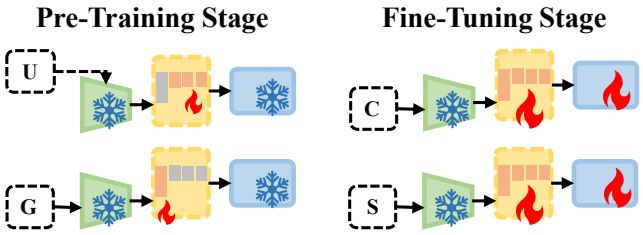

Figure 4: The two-stage training procedure. The pre-training stage focuses solely on the MFAM, while the fine-tuning stage also refines the LLM. The datasets are: unlabeled images (U), generic image-text pairs (G), aesthetic image-comment pairs (C), and aesthetic image-score pairs (S).

et al. 2012), AADB (Kong et al. 2016), EVA (Kang et al. 2020), ICAA, PCCD (Chang et al. 2017), pexels (Pfister et al. 2021), SPAQ (Fang et al. 2020) and TAD66K (He et al. 2022). To refine the learning process, augmentations on quality and color are designed to optimize $\mathrm{H_v^{middle}}$ and $\mathrm{H_v^{low}}$, while those on topics and composition benefit $\mathrm{H_v^{middle}}$ and $\mathrm{H_v^{high}}$. **Generic Pre-Training** focuses on training the fully connection to align $\mathrm{H_v^{thematic}}$, considering the value of generic knowledge for IAA (Ke et al. 2023). The training data comprises a 558K subset of LAION-CC-SBU (Schuhmann et al. 2022; Changpinyo et al. 2021; Saleh and Elgammal 2015) and ShareGPT4V (Chen et al. 2023).

The fine-tuning stage consists of two task-specific processes that fine-tune MFAM and LLM concurrently. **Aesthetic Commenting Fine-Tuning** uses the AVA-Captions dataset (Ghosal et al. 2019) to address AC task. **Aesthetic Scoring and PIAA Fine-Tuning** follows the aesthetic commenting fine-tuning, based on the insight from VILA that mastering AC can bolster effectiveness in AS. We use the AVA dataset for AS and the FLICKR-AES dataset for PIAA.

Having progressed through the two training stages, we are thrilled to find that CALM exhibits a strong aesthetic insight, primarily in the ability to accomplish some zero-shot activities such as giving aesthetic suggestions and conducting in-context PIAA. We highlight an examples of this in Fig. 3 and share more detailed experimental results later.

4 Experiments

4.1 Experimental Settings

Datasets. The AVA dataset comprises over 250,000 images with scores rated by users on the DPChallenge website. We used the official split, designating 19,928 images as the test set and the remainder for training. The AVA-Captions dataset contains approximately 230,000 images, each with an average of 5 user comments. To prevent data leakage, images from the AVA test set are excluded from AVA-Captions training, resulting in 210,000 images for training and 9,361 for testing. FLICKR-AES includes 35,263 images rated by 173 annotators in the training set and 4,737 images evaluated by 37 annotators in the test set, along with user identifications. Additionally, during the pre-training stage, around 460,000 unlabeled images and approximately 660,000 generic image-text pairs were utilized. Notably, we can further expand the unlabeled images as needed.

Implementation Details. The input resolution for $g(\cdot)$ is 224, and $N_v = 256$ visual tokens are processed, each with a dimension of $d_v = 1024$. In subsequent experiments, we set N_{low}, N_{middle} and N_{high} to 32. The dimension of language tokens is $d_l = 4096$. Training was conducted on eight 80GB A100 GPUs, utilizing the Adam optimizer (Kingma and Ba 2014). The peak learning rate was set to 1e-3 for the pre-training stage, and 2.5e-5 and 7e-5 for the two processes in the fine-tuning stage, respectively. Both stages commenced with a linear warm-up, followed by a cosine annealing schedule (Loshchilov and Hutter 2016), with durations of 5 hours and 16.5 hours, respectively.

Evaluation Metrics. For AS, we use Spearman Rank-order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) as metrics. SRCC and PLCC measure the ranking accuracy and linear correlation between the predictions and the ground truth, respectively. For AC, we employ BLEU, ROUGE, CIDEr, and METEOR. BLEU and ROUGE focus on the precision of generated words. CIDEr underscores semantic alignment. METEOR accounts for both semantic and structural similarity. For PIAA, SRCC is adopted again as the primary metric.

Aesthetic Data Extension. CALM strictly followed the procedure outlined in Sec. 3.4 for a fair comparison. Be-

Models	CNN-based models			Trans-based models			CLIP-based models			IAA-adapted MLLMs			IAA-unadapted MLLMs				Ours		
	TANet	ResNext	POC	DAT	MaxViT	EAT	Comm	AesCLIP	CSKD	VILA	Q-Align	UNIAA	LLaVA1.5	BLIP2	miniGPT4	GPT-4v	CALM	CALM-E	CALM-E
Reso.	224	512	640	224	512	224	224	224	224	224	448	336	336	224	224	-	224	224	336
Params	40M	43M	1.9B	87M	31M	87M	-	-	-	383M	8.2B	6.9B	6.9B	2.6B	7.5B	-	7.1B	7.1B	7.1B
FLOPs	-	-	-	240G	120G	140G	-	-	-	-	-	359G	359G	125G	550G	-	770G	770G	878G
PLCC↑	0.765	0.781	0.795	0.739	0.745	0.770	0.740	0.779	0.779	0.774	0.823	0.838	0.083	0.145	0.087	0.412	**0.829**	0.836	**0.852**
SRCC↑	0.758	0.780	0.794	0.738	0.708	0.759	0.734	0.771	0.770	0.774	0.819	0.840	0.077	0.141	0.086	0.406	**0.815**	0.823	**0.841**

Table 1: The comparison results on the AS task. The methods with dark blue marks use extra constructed aesthetic QA data.

Models	Reso.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
LLaVA1.5	336	0.130	0.058	0.022	0.008	0.123	0.000	0.095
BLIP2	224	0.205	0.090	0.035	0.013	0.137	0.037	0.045
miniGPT4	224	0.151	0.066	0.024	0.008	0.077	0.000	0.081
GPT-4v	-	0.116	0.053	0.021	0.008	0.110	0.000	0.100
CWS	-	0.535	0.282	0.150	0.074	0.254	0.059	0.107
Yeo	-	0.464	0.238	0.122	0.063	0.262	0.051	-
VILA	224	0.503	0.288	0.170	0.113	0.262	0.076	-
CALM	224	**0.556**	**0.335**	**0.196**	**0.114**	**0.286**	**0.124**	**0.135**
CALM-E	224	**0.577**	**0.348**	0.204	0.121	0.289	0.160	0.130
CALM-E	336	0.558	0.345	**0.211**	**0.132**	**0.295**	**0.167**	**0.140**

Table 2: The comparison results on the AC task.

Models	SRCC↑	Models	SRCC↑
PAM (Ren et al. 2017)	0.520	Wang (Wang et al. 2019)	0.522
PA (Li et al. 2020)	0.543	BLG (Zhu et al. 2020)	0.561
UG (Lv et al. 2021)	0.559	SOA (Zhu et al. 2021)	0.618
TAPP (Li et al. 2022)	0.591	Hou (Hou et al. 2022b)	0.620
MIR (Zhu et al. 2022)	0.621	AFF (Zhu et al. 2023b)	0.628
CALM	0.632	CALM-In	0.612

Table 3: The comparison results on the PIAA task.

sides, we enhanced CALM by using a number of generic and aesthetic question-answer (QA) data during the aesthetic commenting fine-tuning, resulting in an extended version named **CALM-E**. The generic QA dataset includes CC (Lin et al. 2014), GQA (Hudson and Manning 2019), OCR-VQA (Mishra et al. 2019), TextVQA (Singh et al. 2019), and VG (Krishna et al. 2017), while the aesthetic QA dataset is constructed based on their inherent labels. In concrete, an example QA dialogue is crafted for GPT-3.5 to emulate in terms of style and content. Following this, the images and their labels are provided to generate the intended QA dialogues.

4.2 Comparison to Alternative Approaches

Aesthetic Scoring. Tab. 1 presents a comparison with recent benchmark methods on the AVA dataset. The CNN-based models contain TANet, ResNext and POC (Hou et al. 2022a), which have been effective so far. The transformer-based models include DAT, MaxViT and EAT, which resort to ViT for IAA. The CLIP-based models consist of Comm, AesCLIP, and CSKD (Xu et al. 2023), which benefit from language-image pairs. The MLLM-based models are categorized into IAA-adapted MLLMs and IAA-unadapted MLLMs according to whether they are fine-tuned on IAA data. The IAA-adapted MLLMs are composed of VILA, Q-Align, and UNIAA. For the IAA-unadapted MLLMs, we deployed BLIP2, LLaVA-1.5, and miniGPT4 (Zhu et al. 2023a) locally and called GPT-4v remotely. Because GPT-4v often refuses to answer subjective questions, we only evaluate 7992 images for AS and 5131 images for AC.

When using the same aesthetic labeled data as others, CALM achieves a PLCC of 0.829 and an SRCC of 0.815. Compared to VILA, which is the best IAA-adapted MLLM, CALM shows improvements of 0.055 in PLCC and 0.041 in SRCC. Furthermore, despite evidence from EAT indicating that the higher resolution yields better results, CALM performs better at a smaller resolution than POC, suggesting room for improvement with higher resolution. Additionally, after introducing extra constructed data, similar to Q-Align and UNIAA, and using 336x336 images, CALM-E achieves enhancements of 0.023 in PLCC and 0.026 in SRCC. In conclusion, our CALM-E sets a new benchmark in the AS task.

Aesthetic Commenting. Tab. 2 presents a comparative analysis of our method and previous approaches on the AC task. CWS and Yeo integrate CNN features with an LSTM to generate captions, while VILA relies on CoCa to accomplish it. We use a standard question prompt across all models: "Please comment on the current image aesthetically". Experimental results indicate that CALM surpasses all methods in both word prediction and semantic alignment. As expected, the inclusion of constructed QA data and the increase in input resolutions enhance the model's capacity for aesthetic discernment and linguistic articulation.

Personalized Image Aesthetic Assessment. Previous PIAA methods often design an additional network to learn user preferences and guide the IAA backbone to produce personalized scores. In contrast, CALM accomplishes this without any additional network. Notably, we are the first to introduce MLLMs to PIAA. Following the official protocol, we include 10 images per annotator from the test set into the training set and reserve the remaining images for testing. Subsequently, 7 images are used to construct image-score pairs to optimize the prediction of the remaining 3 images. Tab. 3 displays the test outcomes, with CALM modestly surpassing the current best model (CALM 0.632 vs. AFF 0.628). Besides, the in-context learning version of CALM, i.e., CALM-In, achieves results on par with the leading models, which will be explored in depth in the ablation study.

Qualitative Results. Fig. 5 depicts the aesthetic comments produced by VILA, CALM, and CALM-E. VILA offers basic remarks and omits a deeper analysis. CALM provides more elaborate descriptions but misses finer details. In contrast, CALM-E delivers a detailed analysis and actionable suggestions, showcasing a grasp of aesthetic principles.

4.3 Exploration of Zero-shot Aesthetic Suggesting

To assess zero-shot aesthetic analysis capabilities, we are the first to define the challenging task of aesthetic suggest-

Images			
VILA	*lovely shooting with excellent **color**, great **composition**.*	*maybe could have **cropped a bit more** on top of the birches.*	***color**, **focus** and **saturation** are good, it seems a little **dark**.*
CALM	*Pretty buildings, nice **shadows** on the walls. I love the **colors** and the way you included the tree and the sky.*	*Nice **colors** and **contrast**. I like the way the sun looks, not too overpowering. The **composition** could be improved.*	*Great **colors** and **depth of field**. But if the sun was at a different angle, the colors would not be so washed out.*
CALM-E	*The image is quite captivating. The vibrant **colors** of the buildings and the surroundings really draw the viewer in. The **composition** of the image is well-balanced, with the buildings positioned in a way that creates a visually appealing scene. The **contrast** between the different **colors** and **textures** adds depth to the image and makes it more visually interesting. The use of natural light also adds to its attractiveness…*	*The image of the sunset is stunning, with the warm colors of the sunset creating a beautiful and serene **atmosphere**. The use of the trees to frame the sunset adds depth and interest to the image, drawing the viewer's eye towards the focal point of the sun. The **composition** could also be further refined to create a stronger sense of depth and dimension, perhaps by adjusting the positioning of the trees or the **angle** of the shot…*	*The image of the flowers is appealing, with vibrant **colors** and a nice **depth of field**. The **composition** is well-balanced, and the colors are pleasing to the eye. But there are a few issues that could be improved. While generally good **composition**, there is a lack of a clear **focal point** in the image. Then, the image could benefit from post-processing of **colors** and **contrast**, as well as to **sharpen the details** of the flowers…*

Figure 5: Qualitative comparison of aesthetic commenting. The red comments are correct, while the green ones are wrong.

Types of Degradations	Gaussian or Salt-and-pepper Noise	Motion or Defocus Blur	Brightness Reduction	Brightness Increase	Color Saturation Reduction	Cropping	Avg.
VILA	0.18	0.42	0.48	0.38	0.17	0.88	0.42
LLaVA1.5	0.42	0.54	0.29	0.07	0.47	**1.00**	0.47
GPT-4v	0.63	0.75	0.65	0.43	0.32	0.80	0.60
qwen-vl	0.08	0.43	0.22	0.09	0.16	0.84	0.30
sparkmulti3	0.07	0.39	0.32	0.09	0.21	0.98	0.34
cogvlm	0.12	0.44	0.26	0.04	0.25	0.89	0.33
glm-4v	0.01	0.26	0.22	0.00	0.17	0.98	0.27
CALM	0.40	0.44	0.69	0.37	0.18	0.91	0.50
CALM-E	**0.76**	**0.82**	**0.89**	**0.63**	**0.64**	**1.00**	**0.79**

Table 4: Comparative accuracy of aesthetic suggesting.

ing. This task requires models to provide suggestions for enhancing the aesthetic value of input images. However, due to the numerous factors that influence aesthetics, evaluating the quality of generated suggestions is challenging. Therefore, we artificially impose a severe degradation on each image to prompt the model to suggest improvements for this degradation. Naturally, the model is permitted to include additional suggestions, as these may also impact the overall aesthetics.

For testing, we curated 100 high-quality images from the PCCD dataset and subjected them to six drastic degradations. These degradations include adding Gaussian or salt-and-pepper noise, applying motion or defocus blur, reducing brightness, increasing brightness, reducing color saturation, and cropping the image. We utilize the instructions shown in Fig. 3 to query the MLLMs, and evaluate their responses via GPT-3.5. Specifically, because VILA can only provide a few simple words, we assess its correctness based on whether these words contain the expected degraded attributes.

Quantitative Results. Tab. 4 displays the accuracy of aesthetic suggesting and shows CALM-E achieves the highest accuracy. Trailing closely, GPT4v and CALM exhibit comparable accuracy. While other MLLMs may provide good suggestions for high-level degradations, they fall short in offering advice on image quality and color. Besides, CALM-E significantly outperforms CALM, confirming that ample generic and aesthetic data can enhance aesthetic insight.

Qualitative Results. Fig. 6 shows the generated suggestions for two degraded images and a real image from AVA-Captions. Since the real image lacks definitive answers, we include its annotated score and comment for reference. Each model is given the identical query: "Please suggest some aesthetic improvements to this image". The outcomes reveal that LLaVA-1.5 occasionally hallucinates, and GPT-4v often misses the point. In contrast, CALM-E accurately identifies issues and articulates precise suggestions.

thematic-level	high-level	middle-level	low-level	PLCC↑	SRCC↑
✓				0.768	0.759
✓	✓	✓		0.786	0.778
✓		✓	✓	0.786	0.773
✓	✓	✓	✓	**0.829**	**0.815**

Table 5: Effects of Qformers at different levels in MFAM.

4.4 Ablation Study

Does the MFAM help? To investigate this, we conducted trials by maintaining different Qformers within the MFAM and compared their effects on the AS task. Tab. 5 presents the comparative results. Our baseline maintained only the thematic-level projection, closely resembling LLaVA-1.5. However, with the addition of each layer of Qformers, a notable improvement was observed——a boost by 0.061 in PLCC and 0.056 in SRCC. This clearly demonstrates the necessity of aligning features across all three levels.

How many aesthetic queries in Qformers are optimal? Intuitively, the more the aesthetic queries, the higher the accuracy and computational cost will be. To explore the trade-off between effectiveness and performance, we conducted an ablation study on the AC task. As shown in Fig. 6, with #queries increasing, the effect improves rapidly at first and then stabilizes when the number reaches 32. Besides, due to the attention operations in Qformers and the LLM, the overall computational cost increases quadratically, becoming noticeable if #queries is large. Therefore, we opted for 32 queries per Qformer. Of course, having more queries per Qformer may be better if computing resources are sufficient.

Is the text-guided self-supervised learning useful? To answer this, we removed the self-supervised pre-training from the standard training process for AS task. Evaluation results are shown in the first column of Tab. 7. Compared with the original CALM, CALM without self-supervised learning underperforms by 0.047 in PLCC and by 0.041 in SRCC. Although MFAM structurally facilitates multi-scale feature extraction, the absence of text-guided self-

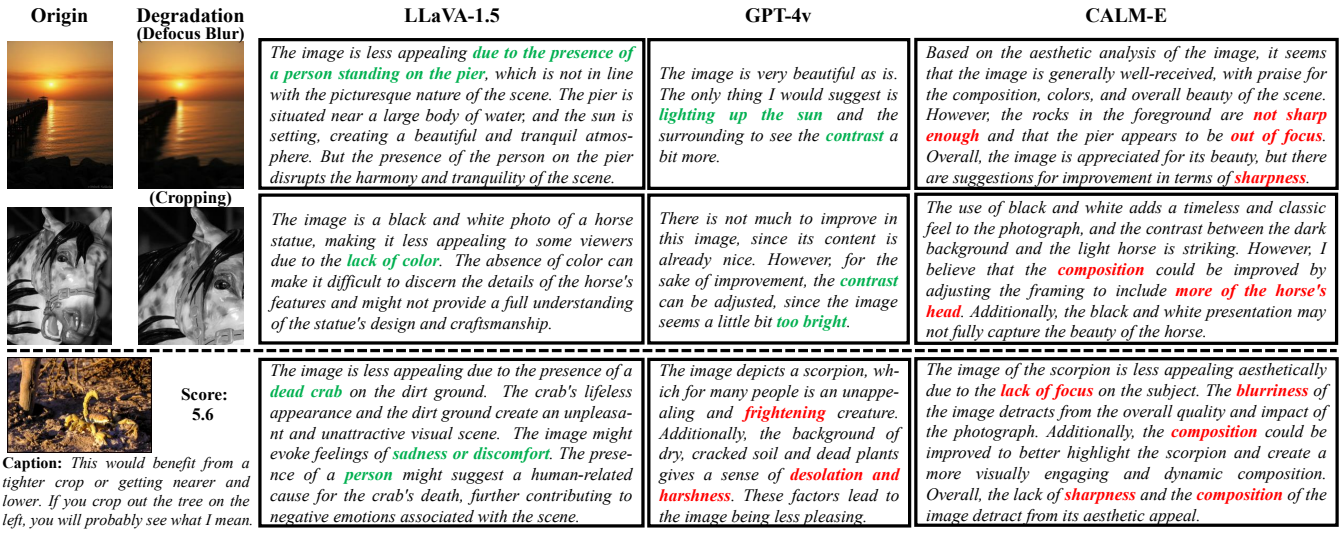

Figure 6: Qualitative comparison of aesthetic suggesting. The red comments are correct, while the green ones are wrong.

#queries	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	METEOR
4	0.509	0.290	0.157	0.080	0.264	0.094	0.123
8	0.537	0.321	0.183	0.099	0.277	0.112	0.128
16	0.544	0.327	0.188	0.103	0.284	0.116	0.129
32	**0.556**	0.335	0.196	0.114	0.286	0.124	**0.135**
64	0.551	**0.336**	**0.199**	**0.118**	**0.289**	**0.142**	0.132

Table 6: Effects vary with #queries in Qformer on AC task.

quality		✓			✓	✓		✓
color			✓		✓		✓	✓
subject				✓		✓	✓	✓
PLCC↑	0.782	0.789	0.789	0.716	0.807	0.788	0.788	**0.829**
SRCC↑	0.774	0.776	0.779	0.780	0.793	0.776	0.775	**0.815**

Table 7: Comparison of using different data augmentation types in the text-guided self-supervised learning on AS task.

supervised learning significantly impairs the overall effects.

Is every type of data augmentation necessary? To delve deeper into the role of each data augmentation type for self-supervised learning, we categorize them into three groups: quality (noise, compression, blur, pixelation), color (brightness, saturation, contrast), and subject (blurring or masking foreground, cropping objects). We then conducted trials on AS task using various combinations of these augmentations. Tab. 7 reveals nuanced insights into the impact of each augmentation type. The findings suggest that solely leveraging quality or color augmentations yields modest improvements, and relying exclusively on subject augmentations appears to provide no significant benefit, likely due to the visual encoder's inherent high-level reasoning capacities. However, a synergy is observed when subject augmentations are introduced alongside quality and color augmentations. This integration not only contributes to improved outcomes but also ensures the preservation of high-level information.

How well is in-context learning suited for PIAA? In-context learning involves providing a model with QA examples of similar questions before asking a specific question, thereby enabling the model to answer such questions. Theoretically, it is ideal for solving the PIAA task, as it allows the model to infer a user's implicit aesthetic tastes from the QA examples. To test this hypothesis, instead of adding 10 images per annotator from the test set to the training set, we randomly select 5 images per annotator to construct the in-context instructions shown in Fig. 3 for use during the test period. Note that CALM-In still requires training on the FLICKR-AES training set; otherwise, it does not perform well on the in-context learning-based PIAA task. Tab. 3 demonstrates that, despite not acquiring the annotators' preferences in advance, CALM-In can elicit user preferences and achieve outcomes comparable to some latest methods.

Are there any limitations? Firstly, Tab. 1 indicates that increasing image resolution can enhance the effect of the AS task. However, our exploration in this aspect is limited, as our visual encoder, the pre-trained ViT-L/14, cannot accommodate varying resolution inputs as flexibly as CNNs. Secondly, the computational burden of our approach is somewhat high, which may limit its applicability scenarios. We intend to address these identified shortcomings in the future.

5 Conclusion

Our study presents CALM, an advanced comprehensive aesthetic large language model. To ensure the extraction of multi-scale aesthetic features both structurally and functionally, we propose the multi-scale text-guided self-supervised learning. Additionally, the instruct-tuning technique is developed to enable CALM to perform multiple aesthetic tasks. Extensive testing reveals that CALM outperforms the current leading approaches across all IAA tasks, solidifying its dominance in the field of IAA. Furthermore, its remarkable zero-shot capabilities in in-context learning PIAA and offering aesthetic suggestions are fully exploited.

References

Chang, K.-Y.; Lu, K.-H.; Chen, C.-S.; and et al. 2017. Aesthetic critiques generation for photos. In *Proceedings of the IEEE international conference on computer vision*, 3514–3523.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, Q.; Zhang, W.; Zhou, N.; Lei, P.; Xu, Y.; Zheng, Y.; and Fan, J. 2020. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14114–14123.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.

Ghosal, K.; Rana, A.; Smolic, A.; and et al. 2019. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

He, S.; Ming, A.; Li, Y.; Sun, J.; Zheng, S.; and Ma, H. 2023a. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21838–21847.

He, S.; Ming, A.; Zheng, S.; Zhong, H.; and Ma, H. 2023b. EAT: An Enhancer for Aesthetics-Oriented Transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1023–1032.

He, S.; Zhang, Y.; Xie, R.; Jiang, D.; and Ming, A. 2022. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 942–948.

Hou, J.; Ding, H.; Lin, W.; Liu, W.; and Fang, Y. 2022a. Distilling knowledge from object classification to aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7386–7402.

Hou, J.; Lin, W.; Yue, G.; Liu, W.; and Zhao, B. 2022b. Interaction-matrix based personalized image aesthetics assessment. *IEEE Transactions on Multimedia*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jin, X.; Wu, L.; Zhao, G.; Li, X.; Zhang, X.; Ge, S.; Zou, D.; Zhou, B.; and Zhou, X. 2019. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM international conference on multimedia*, 311–319.

Kang, C.; Valenzise, G.; Dufaux, F.; and et al. 2020. Eva: An explainable visual aesthetics dataset. In *Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends*, 5–13.

Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; and Yang, F. 2023. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10041–10051.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 662–679. Springer.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, L.; Zhu, H.; Zhao, S.; Ding, G.; and Lin, W. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29: 3898–3910.

Li, Y.; Yang, Y.; Li, H.; Chen, H.; Xu, L.; Li, L.; Li, Y.; and Guo, Y. 2022. Transductive aesthetic preference propagation for personalized image aesthetics assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*, 896–904.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lv, P.; Fan, J.; Nie, X.; Dong, W.; Jiang, X.; Zhou, B.; Xu, M.; and Xu, C. 2021. User-guided personalized image aesthetic assessment based on deep reinforcement learning. *IEEE Transactions on Multimedia*.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Murray, N.; Marchesotti, L.; Perronnin, F.; and et al. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415. IEEE.

Niu, Y.; Chen, S.; Song, B.; Chen, Z.; and Liu, W. 2022. Comment-guided semantics-aware image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1487–1492.

Pfister, J.; Kobs, K.; Hotho, A.; and et al. 2021. Self-supervised multi-task pretraining improves image aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 816–825.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, J.; Shen, X.; Lin, Z.; Mech, R.; and Foran, D. J. 2017. Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision*, 638–647.

Saleh, B.; and Elgammal, A. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sheng, K.; Dong, W.; Chai, M.; Wang, G.; Zhou, P.; Huang, F.; Hu, B.-G.; Ji, R.; and Ma, C. 2020. Revisiting image aesthetic assessment via self-supervised feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5709–5716.

Sheng, X.; Li, L.; Chen, P.; Wu, J.; Dong, W.; Yang, Y.; Xu, L.; Li, Y.; and Shi, G. 2023. AesCLIP: Multi-Attribute Contrastive Learning for Image Aesthetics Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1117–1126.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, 459–479. Springer.

Wang, W.; Su, J.; Li, L.; Xu, X.; and Luo, J. 2019. Meta-learning perspective for personalized image aesthetics assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, 1875–1879. IEEE.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4794–4803.

Xu, L.; Xu, J.; Yang, Y.; Huang, Y.; Xie, Y.; and Li, Y. 2023. CLIP Brings Better Features to Visual Aesthetics Learners. *arXiv preprint arXiv:2307.15640*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.

Yeo, Y.-Y.; See, J.; Wong, L.-K.; and Goh, H.-N. 2021. Generating aesthetic based critique for photographs. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2523–2527. IEEE.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhou, Z.; Wang, Q.; Lin, B.; Su, Y.; Chen, R.; Tao, X.; Zheng, A.; Yuan, L.; Wan, P.; and Zhang, D. 2024. UNIAA: A Unified Multi-modal Image Aesthetic Assessment Baseline and Benchmark. *arXiv preprint arXiv:2404.09619*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, H.; Li, L.; Wu, J.; Zhao, S.; Ding, G.; and Shi, G. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 52(3): 1798–1811.

Zhu, H.; Shao, Z.; Zhou, Y.; Wang, G.; Chen, P.; and Li, L. 2023b. Personalized Image Aesthetics Assessment with Attribute-guided Fine-grained Feature Representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6794–6802.

Zhu, H.; Zhou, Y.; Li, L.; Li, Y.; and Guo, Y. 2021. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia*.

Zhu, H.; Zhou, Y.; Shao, Z.; Du, W.; Wang, G.; and Li, Q. 2022. Personalized Image Aesthetics Assessment via Multi-Attribute Interactive Reasoning. *Mathematics*, 10(22): 4181.