UniQA: Unified Vision-Language Pre-training for Image Quality and Aesthetic Assessment

Hantao Zhou, Longxiang Tang, Rui Yang, Guanyi Qin, Yan Zhang, Yutao Liu, Xiu Li, Runze Hu,
Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Image Quality Assessment (IQA) and Image Aesthetic Assessment (IAA) aim to simulate human subjective perception of image visual quality and aesthetic appeal. Despite distinct learning objectives, they have underlying interconnectedness due to consistent human assessment perception. In this paper, we propose Unified vision-language pre-training of Quality and Aesthetics (UniQA), to extract useful and common representations from two tasks, thereby benefiting them simultaneously. However, the lack of text in the IQA datasets and the textual noise in the IAA datasets pose severe challenges for multimodal pre-training. To address this, we (1) utilize multimodal large language models (MLLMs) to generate high-quality text descriptions; (2) use the generated text for IAA as metadata to purify noisy IAA data. To effectively adapt the pre-trained UniQA to downstream tasks, we further propose a lightweight adapter that utilizes versatile cues to fully exploit the extensive knowledge of the pre-trained model. UniQA demonstrates high competitiveness in various image assessment tasks, including classical IQA and IAA tasks, few-label IQA, and other downstream tasks, showing promise as a foundational assessment model. Codes are available at https://github.com/zht8506/UniQA.**

***Index Terms*—Image Quality Assessment, Image Aesthetic Assessment, Vision-Language Pre-training, Multimodal Large Language Models.**

I. Introduction

Image Quality Assessment (IQA)[1] and Image Aesthetic Assessment (IAA) aim to measure the perceived quality and beauty of an image. They find broad applications in many scenarios, such as guiding individuals in image photography and editing, and serving as tools for image dehazing model [1]. Consequently, huge efforts [2], [3], [4], [5] have been devoted to establishing effective IQA and IAA models.

IQA and IAA concentrate on distinct aspects of image assessment, with IQA primarily focusing on the distortion level of the image, while IAA is oriented towards evaluating the aesthetic appeal of the image. Despite their differences, IQA and IAA have underlying commonality: **simulating human subjective perceptions of images.** Specifically, in human subjective image evaluation, quality and aesthetics exhibit a mutual influence [6], such that high-quality images are more likely to possess a higher aesthetic appeal compared to their low-quality counterparts. Thus, the learning process for both tasks not only acquires features unique to themselves but also involves the learning of task-agnostic common representations [7]. This commonality sparks an idea:

> *Can we develop a foundational model with robust visual assessment perceptions consistent with human to benefit both IQA and IAA tasks?*

Several existing works [3], [8], [9] have explored the relationship between IQA and IAA tasks from different perspectives. For instance, some works (*e.g.*, MUSIQ [3]) can be applied to IQA and IAA tasks indiscriminately, but they cannot exploit beneficial representations from another task. Q-Align [10] and DSINet [11] also find the similarities of two tasks and attempt to tackle them with unified architecture and training. However, they typically unify datasets of two tasks for regression training directly, which cannot explicitly learn the task-shared representations, restricting the extraction of mutual benefits. TQ4AQ [9] develops a quality-assisted image aesthetic quality assessment method that utilizes quality information to improve the IAA task, but only considers the impact of IQA on IAA. In this paper, we propose the **Uni**fied pre-training of **Q**uality and **A**esthetics (UniQA) to extract mutually beneficial and effective representations for both tasks. Then, the pre-trained UniQA can be flexibly applied to IQA and IAA datasets.

To achieve unified pre-training, a straightforward solution involves consolidating all IQA and IAA datasets and then training the model to regress towards the mean opinion scores (MOS) annotated by humans. However, existing datasets show variations in perceptual scales due to differences in subjective testing methodologies [12]. As a result, this training strategy makes the model develop a score bias toward larger scale datasets. Moreover, it may not effectively capture the unique characteristics of IQA and IAA, as the MOS labels cannot be explicitly interpreted. To this end, we propose to use **text descriptions** as a bridge to integrate the two tasks, leveraging the rich and fine-grained semantics inherent in text to provide more auxiliary information.

To achieve unified pre-training, we need to construct image-text data for IQA and IAA tasks. However, existing IQA

Hantao Zhou, Longxiang Tang, Rui Yang, Guanyi Qin, Xiu Li are with Tsinghua Shenzhen International Graduate School, Tsinghua University, ShenZhen, 518055, China (e-mail: hantaozh@outlook.com, lloong.x@gmail.com, r-yang20@mails.tsinghua.edu.cn, qgy21@mails.tsinghua.edu.cn, li.xiu@sz.tsinghua.edu.cn).

Runze Hu is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100086, China (e-mail: hrzlpk2015@gmail.com).

Yan Zhang is with Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University, 361005, China (e-mail: bzhy986@gmail.com)

Yutao Liu is with the School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: liuyutao@ouc.edu.cn)

Guangtao Zhai is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaiguangtao@gmail.com).

[1] The IQA in this work refers to the no-reference IQA.

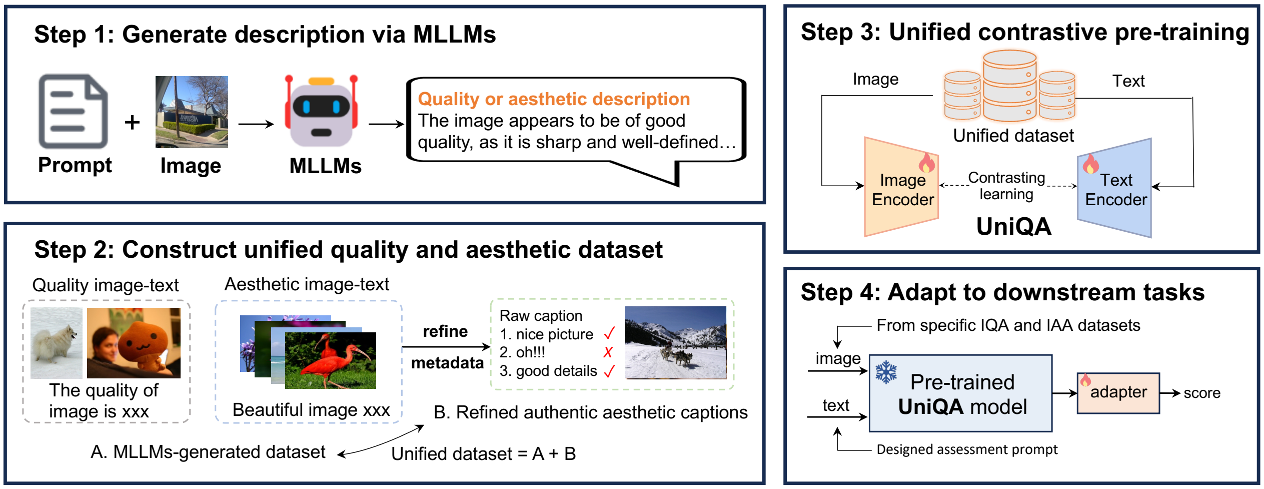

Step 1: Generate description via MLLMs
Prompt
+
Image
MLLMs
Quality or aesthetic description
The image appears to be of good quality, as it is sharp and well-defined…
Step 2: Construct unified quality and aesthetic dataset
Quality image-text
Aesthetic image-text
The quality of image is xxx
Beautiful image xxx
refine
metadata
Raw caption
1. nice picture ✓
2. oh!!! ✗
3. good details ✓
B. Refined authentic aesthetic captions
A. MLLMs-generated dataset
Unified dataset = A + B
Step 3: Unified contrastive pre-training
Image
Text
Unified dataset
Image Encoder
Contrasting learning
Text Encoder
UniQA
Step 4: Adapt to downstream tasks
From specific IQA and IAA datasets
image
text
Pre-trained UniQA model
adapter
score
Designed assessment prompt

Fig. 1: The overview of our method. We leverage MLLMs to generate quality- and aesthetics-related descriptions (Step 1) and utilize the generated data to refine authentic noisy data (Step 2). We conduct unified pre-training to obtain UniQA (Step 3), which can be flexibly applied to both IQA and IAA tasks with a lightweight adapter (Step 4).

datasets [13], [14], [15], [16], [17], [18] typically have images only and lack text descriptions. While IAA dataset [19] include text data provided by humans, they often contain considerable textual noise irrelevant for aesthetic assessment. Therefore, a top priority is determining how to acquire high-quality image-text data for both tasks. Recently, multimodal large language models (MLLMs) [20], [21], [22], [23] have demonstrated outstanding capabilities in image understanding, which can generate reasonable responses based on images and user instructions. Inspired by this, we propose utilizing MLLMs with tailored prompts to generate quality- and aesthetics-related descriptions for the IQA and IAA datasets (Step 1 of Fig. 1). As observed in Fig. 2, this approach provides a comprehensive and precise depiction of image quality and aesthetics. Furthermore, we utilize these generated high-quality aesthetic descriptions as metadata to refine the raw aesthetic caption dataset (Step 2 of Fig. 1). Finally, we unify the IQA and IAA datasets to conduct multimodal pre-training (Step 3 of Fig. 1) to obtain UniQA.

To effectively adapt the pre-trained UniQA to downstream tasks, we propose a lightweight adapter, namely the Multi-Cue Integration Adapter (Step 4 of Fig. 1). This adapter uses versatile cues related to image assessment to prompt the pre-trained UniQA, adeptly extracting useful knowledge and comprehensively assessing the image. With much fewer tunable parameters compared to previous IQA and IAA models, our model outperforms them on both tasks. More surprisingly, our method achieves impressive results on few-label IQA, *e.g.*, achieving the SRCC values of 0.828 (vs. 0.760 on CLIVE of SOTA method GRepQ [24]). UniQA also generalizes well to various downstream image assessment tasks, highlighting its strong foundation and generalization capabilities.

Our contributions can be summarized as follows:

- With the assistance of MLLMs, we construct a high-quality image-text dataset about image quality and aesthetics. Thro-ugh pre-training on this dataset, we develop UniQA, which effectively learns a general perception of image assessment, promoting the effective and efficient learning of both IQA and IAA tasks.
- We propose a novel Multi-Cue Integration Adapter, which integrates various assessment-related cues to fully exploit the extensive knowledge of the pre-trained model with minimal additional parameters.
- Extensive experiments show that our method achieves impressive performance across classical IQA and IAA tasks, few-label IQA and other downstream tasks, showing promise as a foundational assessment model.

II. Related Work

A. Image Quality Assessment

The rapid development of deep learning has sparked significant interest in their application for IQA. Many researchers utilize CNN to solve the IQA problem with various effective techniques, including multi-level feature aggregation [25], adaptive quality prediction [2], and patch-to-picture learning [18], and unsupervised learning [26]. Recently, transformer-based IQA methods [3], [27], [28], [29] show promising results in the IQA field, which can compensate for the non-local representation ability of CNN. Despite these impressive breakthroughs, these methods often transfer models pre-trained on classification datasets, such as ImageNet [30], to IQA tasks, which may be suboptimal [31]. Q-Align [10] attempts to jointly perform IQA and IAA tasks, but it uses a huge language model and does not explicitly extract features of the two tasks through pre-training. Our method can learn more effective representations through joint pre-training on quality-aesthetics image-text data, benefiting IQA tasks.

B. Image Aesthetic Assessment

Image Aesthetic Assessment (IAA) aims to measure the aesthetic quality of images. With the advent of deep learning, IAA methods have evolved from hand-crafted feature extraction [32], [33], [34] to end-to-end feature learning, marking significant advancements in the IAA domain. Various techniques have been developed to boost IAA task, including local and global feature integration [35], [36], [37], [38], [39], graph network [40], [41], knowledge distillation [42] and theme-aware learning [43], [4]. Recently, there has been an emergence of multimodal IAA methods [44], [45], [46], [47] that incorporate text as auxiliary supervision. However, these methods necessitate the use of text during inference, limiting their flexible application since text is often not easily available. Our method overcomes this limitation by conducting vision-language pre-training firstly to learn effective representation. The pre-trained model can be flexibly applied to the IAA field using only images.

C. Vision-Language Models

Vision-Language Models (VLMs) [48], [49], [50], [51] introduce the contrastive learning strategy to acquire image-text correspondences from large-scale image-text pairs. VLMs have exhibited promising results across multiple tasks, including IQA [52], [53] and IAA [54], [55]. Recently, the Multimodal Large Language Models (MLLMs) have garnered increasing research interest, exhibiting remarkable prowess in comprehending image content and reasoning through complex instructions [21], [56], [23], [57], [20]. Most existing MLLMs achieve this by integrating image features with LLM tokens, subsequently fine-tuning the LLM via multimodal instruction tuning. During inference, MLLMs can reason with given images and user instructions, generating text responses by leveraging world knowledge learned during pre-training.

III. UniQA: MLLMs-assisted Unified Pre-training

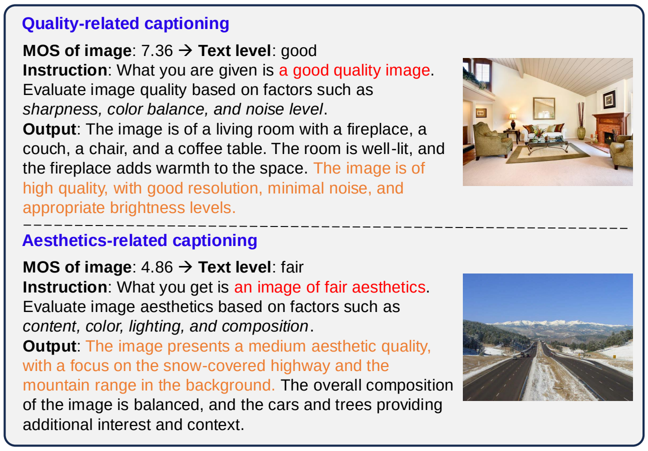
In this section, we first present some preliminaries of related models. We then describe the process of constructing a unified image-text dataset about quality and aesthetics, with the assistance of MLLMs (Section III-B and III-C). We use this dataset for vision-language pre-training to construct UniQA.

A. Preliminaries

Vision-language pre-training aims to achieve comprehensive cross-modality understanding by training on web-scale image-text datasets. Benefiting from this large-scale pre-training, CLIP, a prominent VLM, has demonstrated great promise to assist a broad scope of vision tasks. Specifically, CLIP comprises an image encoder f and a text encoder g, both jointly trained to establish a shared latent space for image and text through contrastive learning.

Given a batch of N paired images and texts $\{x_I^i, x_T^i\}_{i=1}^N$, CLIP extracts image features $\boldsymbol{I} = \{f(x_I^i)\}_{i=1}^N$ and text features $\boldsymbol{T} = \{g(x_T^i)\}_{i=1}^N$ with corresponding encoders. During pre-training, CLIP seeks to maximize the cosine similarity of paired image and text features, while minimizing the similarity of unmatched pairs. The contrastive learning objective can be formulated as:

$$\begin{aligned} \mathcal{L}_{\text{image}} &= -\mathbb{E}_{I_i \sim \boldsymbol{I}} \left[\log \frac{\exp(I_i^\top T_i/\tau)}{\sum_{j=1}^N \exp(I_i^\top T_j/\tau)} \right] \\ \mathcal{L}_{\text{text}} &= -\mathbb{E}_{T_i \sim \boldsymbol{T}} \left[\log \frac{\exp(T_i^\top I_i/\tau)}{\sum_{j=1}^N \exp(T_i^\top I_j/\tau)} \right] \end{aligned} \tag{1}$$



Quality-related captioning

MOS of image: 7.36 → **Text level**: good
Instruction: What you are given is a good quality image. Evaluate image quality based on factors such as *sharpness, color balance, and noise level*.
Output: The image is of a living room with a fireplace, a couch, a chair, and a coffee table. The room is well-lit, and the fireplace adds warmth to the space. The image is of high quality, with good resolution, minimal noise, and appropriate brightness levels.

Aesthetics-related captioning

MOS of image: 4.86 → **Text level**: fair
Instruction: What you get is an image of fair aesthetics. Evaluate image aesthetics based on factors such as *content, color, lighting, and composition*.
Output: The image presents a medium aesthetic quality, with a focus on the snow-covered highway and the mountain range in the background. The overall composition of the image is balanced, and the cars and trees providing additional interest and context.

Fig. 2: Generating quality- and aesthetics-related captions via MLLMs and our MOS-guided task-specific prompts. The red text refers to MOS-based text guidance. The orange text highlights the quality- and aesthetics-related text.

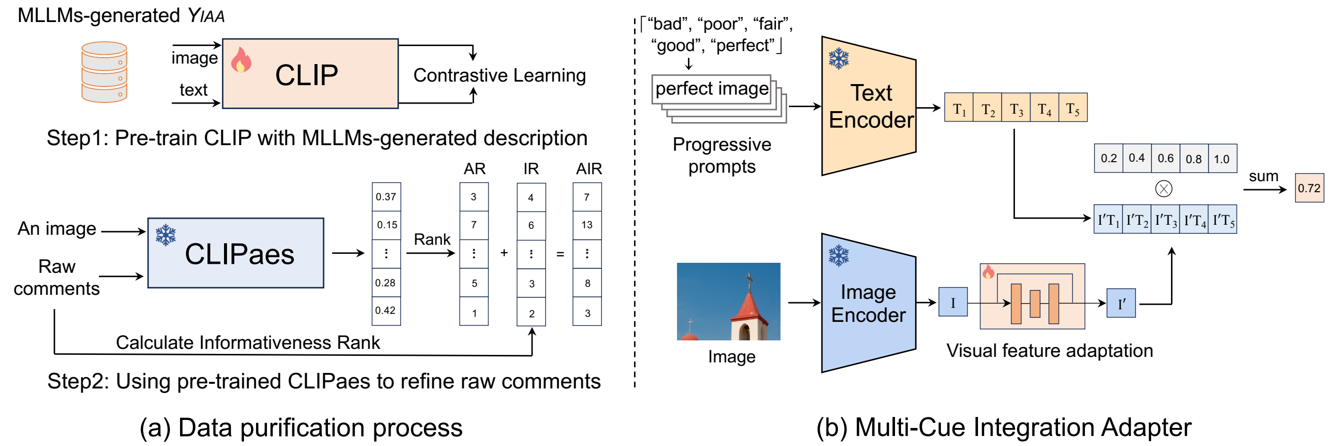
where the I_i and T_i are the i-th features in the batch, and τ is the temperature parameter. The final contrastive learning loss can be obtained by taking the average: $\mathcal{L} = (\mathcal{L}_{\text{image}} + \mathcal{L}_{\text{text}})/2$. With this training strategy, CLIP can generate aligned features for paired image-text samples.

B. Quality- and Aesthetics-related Captioning

In order to achieve vision-language pre-training in the field of image assessment, we need to generate text for IQA and IAA datasets since IQA datasets lack text and IAA datasets contain noisy text. Recently, MLLMs have shown advanced performance, so we can use them to generate high-quality textual data for images. Previous studies [58], [59] have highlighted that it is challenging for MLLMs to directly and accurately perceive the quality and aesthetics of input images, often resulting in positively skewed expressions and strong hallucinations. Thus, to obtain correct and detailed descriptions about quality and aesthetics, as shown in Fig. 2, we design **MOS-guided task-specific prompts** to instruct MLLMs:

$$Y_t \sim M_T(x_I, P_t|G). \tag{2}$$

where M_T denotes the used MLLM, G is the MOS-based text guidance, P_t is the task-specific prompt, Y_t represents the generate caption. To obtain G, we divide images into 5 levels based on MOS, *i.e.*, $\{$`bad`, `poor`, `fair`, `good`, `perfect`$\}$ [14], [53]. If an image's MOS ranks in the top 20% of the score range, its level is assigned to `perfect`. This approach harmonizes IQA and IAA datasets with different MOS scales, alleviating the MOS biases of different datasets [12]. Additionally, P_t is customized for IQA (P_{IQA}) and IAA (P_{IAA}) tasks, respectively. As shown in Fig. 2, P_{IQA} involves *sharpness, color balance, and noise level* [60], while P_{IAA} includes *content, color, lighting, and composition* [61]. With these designs, M_T is guided towards image assessment and we can obtain generated caption datasets Y_{IQA} and Y_{IAA} for IQA and IAA tasks, respectively. For simplicity



(a) Data purification process

(b) Multi-Cue Integration Adapter

Fig. 3: (a) Data purification process: we pre-train CLIP using generated aesthetic captions data Y_{IAA} and then use the pre-trained CLIP$_{aes}$ to purify data. (b) The proposed adapter: we employ progressive prompts, {`bad`, `poor`, `fair`, `good`, `perfect`} with “image”, to prompt the frozen UniQA and a lightweight trainable module to adjust visual features.

and cost-effectiveness, we use open-source LLaVA [62] as the captioner. We also experiment with the effects of different MLLMs on model performance (Table X).

C. Data Purification Strategy

In addition to the generated aesthetic captions Y_{IAA}, there are also IAA datasets with captions commented by humans [19], which directly reflect human aesthetic feelings. Incorporating comments from various people can offer a more comprehensive description of image aesthetics. However, it may introduce noise to the data, as individuals may provide comments unrelated to image aesthetics. To address this issue, we propose a novel data purification strategy to refine raw captions in the original dataset. This process is illustrated in Fig. 3(a).

Specifically, we introduce **Aesthetics-relevance and Informativeness Rank (AIR)** to measure the quality of text corresponding to an image. The AIR consists of Aesthetics-relevance Rank (AR) and Informativeness Rank (IR). To obtain AR, we first pre-train a CLIP model with generated aesthetic data Y_{IAA} to get an aesthetics-aware CLIP model, denoted as CLIP$_{\text{aes}}$. Then, we employ it to measure the aesthetics relevance score (s_A) for an image-text pair. Given an image with n captions, AR can be defined as:

$$\text{AR} = \text{Rank}(s_A^1 \cdots s_A^n), \quad s_A^i = \text{CLIP}_{\text{aes}}(x_I, x_T^i), \tag{3}$$

where s_A^i represents the aesthetics relevance score between the i-th caption x_T^i and its corresponding image x_I. Note that AR consists of *long integers* that represent the rank of a caption after sorting by s_A. For an image with n textual captions, IR can be expressed as:

$$\text{IR} = \text{Rank}(s_I^1, \cdots, s_I^n), \quad s_I^i = \text{InfoMeasure}(x_T^i), \tag{4}$$

where $\text{InfoMeasure}(\cdot)$ can output the informativeness score (s_I) of an input sentence. The informativeness score of the text can be measured by various methods, such as sentence length or Shannon entropy [63]. Here we simply use the length of the sentence as the informativeness score and discuss other methods in ablation studies (Table X). As a result, the AIR between an image and n captions is:

$$\text{AIR} = \text{Rank}((\alpha\text{AR}^1 + \beta\text{IR}^1), \cdots, (\alpha\text{AR}^n + \beta\text{IR}^n)), \tag{5}$$

where α and β are used to balance AR and IR to purify the data more flexibly. In implementation, we set them to one for simplicity. We select captions with Top-K ranking AIR to construct a high-quality aesthetic caption dataset, denoted as Y_{IAA}^+. This strategy ensures that the text is both aesthetically relevant and informative, thereby improving the quality and richness of the raw dataset.

D. Unified Vision-Language Pre-training

So far, we have gotten a high-quality image-text dataset about quality and aesthetics, $Y = Y_{IQA} \cup Y_{IAA} \cup Y_{IAA}^+$. Based on it, we pre-train CLIP using Equation 1 to obtain our UniQA. In this way, the model learns general perceptions of image quality and aesthetics, which can provide potent assessment priors and thus can be effectively applied to both IQA and IAA tasks.

IV. Adapting UniQA for IQA and IAA

The pre-trained UniQA contains extensive perception information, which can facilitate downstream assessment tasks in a zero-shot or supervised manner. In this section, we further propose a meticulously designed adapter and prompt ensemble strategy to enhance the model's performance.

A. Multi-Cue Integration Adapter

During pre-training, the model aligns image and assessment-related captions, empowering it with strong comprehension of image quality and aesthetics. With this foundation model, we can slightly adjust the visual features, efficiently adapting it to score-based image assessment tasks. To this end, we introduce a lightweight adapter, namely the **Multi-Cue Integration Adapter**, to adapt visual features and inject rich cues for

TABLE I: RESULTS ON IQA DATASETS. **BLACK** AND **BLUE** NUMBERS IN BOLD REPRESENT THE BEST AND SECOND BEST, RESPECTIVELY. † DENOTES UNFREEZING THE UNIQA BACKBONE FOR FINE-TUNING. HIGHER SRCC AND PLCC IMPLY BETTER PERFORMANCE

	LIVE		TID2013		CSIQ		KADID		CLIVE		KonIQ		SPAQ	
Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [64]	0.436	0.459	0.626	0.571	0.812	0.748	0.528	0.567	0.629	0.629	0.681	0.685	0.809	0.817
WaDIQaM [65]	0.960	0.955	0.835	0.855	0.852	0.844	0.739	0.752	0.682	0.671	0.804	0.807	0.840	0.845
DBCNN [66]	0.968	0.971	0.816	0.865	0.946	0.959	0.851	0.856	0.851	0.869	0.875	0.884	0.911	0.915
MetaIQA [67]	0.960	0.959	0.856	0.868	0.899	0.908	0.762	0.775	0.802	0.835	0.850	0.887	-	-
PaQ-2-PiQ [18]	0.959	0.958	0.862	0.856	0.899	0.902	0.840	0.849	0.844	0.842	0.872	0.885	-	-
HyperIQA [2]	0.962	0.966	0.840	0.858	0.923	0.942	0.852	0.845	0.859	0.882	0.906	0.917	0.911	0.915
TReS [68]	0.969	0.968	0.863	0.883	0.922	0.942	0.859	0.858	0.846	0.877	0.915	0.928	-	-
MUSIQ [3]	0.940	0.911	0.773	0.815	0.871	0.893	0.875	0.872	0.702	0.746	0.916	0.928	0.918	0.921
DACNN [69]	0.978	0.980	0.871	0.889	0.943	0.957	0.905	0.905	0.866	0.884	0.901	0.912	0.915	0.921
DEIQT [27]	0.980	0.982	0.892	0.908	0.946	0.963	0.889	0.887	0.875	0.894	0.921	0.934	0.919	0.923
LIQE [53]	0.970	0.951	-		0.936	0.939	0.930	0.931	**0.904**	0.911	0.919	0.908	-	-
Re-IQA [26]	0.970	0.971	0.804	0.861	0.947	0.960	0.872	0.885	0.840	0.854	0.914	0.923	0.918	0.925
LoDA [28]	0.975	0.979	0.869	0.901	-	-	0.931	0.936	0.876	0.899	0.932	**0.944**	**0.925**	**0.928**
Q-Align [10]	0.913	0.919	-	-	0.915	0.936	0.869	0.927	**0.931**	**0.921**	**0.935**	0.934	-	-
Ours	**0.981**	**0.983**	**0.916**	**0.931**	**0.963**	**0.973**	**0.940**	**0.943**	0.890	0.905	0.933	0.941	**0.924**	**0.928**
Ours†	**0.985**	**0.985**	**0.950**	**0.959**	**0.977**	**0.980**	**0.966**	**0.968**	0.902	**0.914**	**0.940**	**0.948**	**0.925**	**0.929**

fine-tuning downstream tasks. The adapter consists of two key processes: visual feature adaptation and multi-cue integration prediction.

Visual Feature Adaptation. We add a learnable residual module following the image encoder to adjust the visual features so as to adapt to specific assessment datasets. We optimize this module while keeping the image and text backbones frozen, enabling parameter-efficient tuning. The adapter is illustrated in Fig. 3(b). Let I denote the image features extracted from the frozen image encoder, the visual feature adaptation process can be expressed as:

$$I' = \text{Normalize}(\text{Adapter}(I) + I) \tag{6}$$

where the $\text{Adapter}(\cdot)$ consists of two fully connected layers with a ReLU activation in between, and I' represents the adapted visual features.

Multi-cue Integration Prediction. A straightforward approach to incorporating CLIP into image assessment is to use the "`good image`" as an anchor and take the cosine similarity between the text anchor and a given image as the assessment score. However, this method shows two shortcomings: (1) using the absolute value of similarity as score may not be optimal because it only reflects the semantic similarity between images and texts [58], [52]; (2) a single prompt may not fully leverage the extensive knowledge of the pre-trained model. Thus, we propose to utilize versatile cues to comprehensively explore the power of the pre-trained UniQA and convert absolute similarity scores into relative values for weighting.

Specifically, we utilize the prompt template "{level} image" and five text levels (`bad`, `poor`, `fair`, `good`, `perfect`), *i.e.*, "*Multi-cue*", to construct prompts. Next, we calculate the cosine similarity between the normalized text features $\{T_i\}_{i=1}^{5}$ of five prompts and adapted visual features I', and then use the $\text{Softmax}(\cdot)$ to obtain the related value of five image-text correspondence. These related values will weight the predefined score levels to get the final assessment score. The process is described in Fig. 3(b) and can be formulated as follows:

TABLE II: THE ASSESSMENT-ORIENTED ENSEMBLE PROMPTS USED IN ZERO-SHOT AND FEW-LABEL LEARNING

Task	Prompt
CLIVE KonIQ LIVE	{`bad`, `poor`, `fair`, `good`, `perfect`} with `image`
	{`extremely blurry`, `blurry`, `fair`, `sharp`, `extremely sharp`} with `image`
	{`extremely noisy`, `noisy`, `fair`, `noise-free`, `extremely noise-free`} with `image`
	{`extremely low-quality`, `low-quality`, `fair`, `high-quality`, `extremely high-quality`} with `image`
AGIQA-3K	{`bad`, `poor`, `fair`, `good`, `perfect`} with `image` {`bad`, `poor`, `fair`, `good`, `perfect`} with `content`

$$q = \sum_{i=1}^{5} \frac{c_i \exp({I'}^{\top} T_i/\tau)}{\sum_{j=1}^{5} \exp({I'}^{\top} T_i/\tau)}, \tag{7}$$

where $\{c_i\}_{i=1}^{5}$ are scores of text levels with progressive values that are set to $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. Note that the scores are learnable parameters and can be dynamically adjusted based on the datasets. τ is the temperature parameter and q is the assessment score of the given image.

B. Assessment-oriented Prompt Ensemble

We introduce an **assessment-oriented prompt ensemble** strategy, which incorporates more image assessment-related prompt groups to derive the final assessment score, thereby achieving a more comprehensive understanding of image quality and aesthetics. For instance, we can use *e.g.*, {`extremely blurry`, `blurry`, `fair`, `sharp`, `extremely sharp`} as another five text levels. Now, the final assessment score q_f is the average of all prompt groups and it can be described as:

$$q_f = \frac{\sum_{i=1}^{m} q_i}{m}, \tag{8}$$

TABLE III: RESULTS ON AVA DATASET

Method	SRCC	PLCC
NIMA [70]	0.612	0.636
MaxViT [71]	0.708	0.745
MUSIQ [3]	0.726	0.738
MLSP [72]	0.756	0.757
TANet [4]	0.758	0.765
MILNet [37]	0.732	0.753
EAT [39]	0.759	0.770
VILA [5]	0.774	0.774
Ours	**0.776**	**0.776**
Ours†	**0.782**	**0.782**

TABLE IV: RESULTS ON AADB DATASET

Method	SRCC	PLCC
NIMA [70]	0.708	0.711
MLSP [72]	0.725	0.726
MUSIQ [3]	0.706	0.712
PA-IAA [73]	0.720	0.728
HIAA [74]	0.739	-
TANet [4]	0.738	0.737
Celona *et al.* [75]	0.757	0.762
TAVAR [43]	0.761	0.763
Ours	**0.786**	**0.787**
Ours†	**0.788**	**0.791**

TABLE V: RESULTS ON BAID DATASET

Method	SRCC	PLCC
NIMA [70]	0.393	0.382
MP$_{ada}$ [76]	0.437	0.425
MLSP [72]	0.441	0.430
BIAA [77]	0.389	0.376
TANet [4]	0.453	0.437
SAAN [78]	0.473	0.467
TSC-Net [79]	0.480	0.479
EAT [39]	**0.486**	0.495
Ours	**0.487**	**0.528**
Ours†	0.484	**0.502**

where m denotes the number of prompt groups. This strategy can more fully utilize the multi-modal understanding capabilities of UniQA and demonstrates non-negligible performance improvements in zero-shot (Table VII) and few-label learning (Table IX). The details of ensemble prompts are represented in Table II. Note that the prompts for AGIQA-3K differ from other IQA datasets, as distortions in AI-generated images is different from those in authentic images. For instance, authentic image distortions may stem from camera shake, whereas AI-generated image distortions typically involve meaningless content and distorted poses. Thus, “content” is used to prompt UniQA for the AGIQA-3K dataset.

V. EXPERIMENTS

A. Datasets

We employ the IQA dataset FLIVE [18] and the IAA dataset AVA [80] for quality- and aesthetics-related captioning, and AVA-Captions [19] to provide authentic aesthetic comments. We evaluate the performance on typical IQA and IAA datasets, including seven IQA datasets and three IAA datasets. We also evaluate the generalization ability of UniQA on three downstream IQA datasets.

IQA Dataset. For the IQA task, four synthetic datasets, including LIVE [13], CSIQ [81], TID2013 [82], KADID [17], and three authentic datasets of CLIVE [14], KonIQ [15], SPAQ [16], are used for performance evaluation. FLIVE [18] is an authentic IQA dataset that contains 39,810 images. We employ three datasets to evaluate the generalization capability of our UniQA, including an AI-generated IQA dataset AGIQA-3K [83], an underwater IQA dataset UWIQA [84] and an enhanced colonoscopy image quality assessment dataset ECIQAD [85]. The details of the used datasets are shown in Table VI.

IAA Dataset. For the IAA task, we conduct experiments on three datasets, including AVA [80], AADB [86] and BAID [78] datasets. AVA comprises 250k images, with the test set of 19,928 images. AADB dataset consists of 10,000 images in total, with 8,500 images for training, 500 images for validation, and 1,000 images for testing. BAID is a large-Scale artistic image aesthetics Assessment dataset, with 60,337 images in tatol. We follow the standard data split and use 53,937 images for training and 6,400 images for testing.

AVA-Captions. AVA-Captions offer multiple human-annotated comments for each AVA image. To avoid potential data leakage, we strictly follow the official data split of AVA, results in a pre-training image-text dataset comprising 234,090 images paired with 3.0 million captions.

TABLE VI: DETAILS OF DIFFERENT IQA AND IAA DATASETS. DIST. NO. INDICATES THE NUMBER OF DISTORTION TYPES

Dataset	Task	Type	Size	Dist. No.
LIVE [13]	IQA	Synthetic	799	5
CSIQ [81]	IQA	Synthetic	866	5
TID2013 [82]	IQA	Synthetic	3,000	24
KADID [17]	IQA	Synthetic	10,125	25
CLIVE [14]	IQA	Authentic	1,162	-
KonIQ [15]	IQA	Authentic	10,073	-
SPAQ [16]	IQA	Authentic	11,000	-
FLIVE [18]	IQA	Authentic	39,810	-
AGIQA-3K [83]	IQA	Authentic	2,982	-
UWIQA [84]	IQA	Authentic	890	-
ECIQA [85]	IQA	Authentic	2400	-
AVA [80]	IAA	Authentic	250,000	-
AADB [86]	IAA	Authentic	10,000	-
BAID [78]	IAA	Authentic	60,337	-

B. Evaluation Criteria

We employ Spearman’s Rank-order Correlation Coefficient (SRCC) and Pearson’s Linear Correlation Coefficient (PLCC) as criteria to measure the performance of IQA and IAA models. They reflect the prediction monotonicity and prediction accuracy of the model, respectively. Both SRCC and PLCC range from 0 to 1. Higher values of SRCC and PLCC indicate better performance. For each IQA dataset, 80% of the images are used for training and the remaining 20% for testing. We repeat this process 10 times to mitigate the performance bias and the medians of SRCC and PLCC are reported. For the IAA datasets, we follow the standard data splits.

C. Implementation Details

In this section, we introduce the details of large-scale assessment-related text data generation and purification, multimodal pre-training to build UniQA, and fine-tuning UniQA for IQA and IAA tasks.

Text data generation and purification. We use LLaVA-1.5-7B [21], [62] as the multimodal large language model (MLLM) for captioning. We generate three captions for each

TABLE VII: SRCC ON THE CROSS DATASETS VALIDATION. * DENOTES USING ENSEMBLE PROMPTS. THE RESULTS OF OTHER METHODS ARE PRE-TRAINED ON FLIVE

Method	CLIVE	KonIQ	AGIQA-3K
DBCNN [66]	0.724	0.716	0.645
PaQ-2-PiQ [18]	0.738	0.755	0.502
HyperIQA [2]	0.735	0.758	0.629
TReS [68]	0.740	0.713	0.646
DEIQT [27]	**0.781**	**0.733**	-
CLIP* [48]	0.746	0.592	0.646
Ours	0.638	0.667	**0.744**
Ours*	**0.790**	**0.806**	**0.752**

IQA image and one caption for each IAA image, resulting 119,421 generated IQA captions and 234,090 IAA captions. We use these high-quality generated IAA data as metadata to purify the raw AVA-Captions dataset. We set $K = 4$ to refine the AVA-Captions dataset. In the end, we obtain 273,897 images, and paired with 1,240,915 comments. Each image has at least three captions.

Multimodal pre-training. We adopt CLIP-B/16 [48] as the vision-language model (VLM) for pre-training. We follows the same pre-training strategy as CLIP to train our UniQA. We use the Adam optimizer [87] with a learning rate of 5e-6 and weight decay of 0.2. The training is performed for 5 epochs with a batch size of 960, and it is resource-efficient, taking *less than an hour per run* on four A100 GPUs.

Fine-tuning UniQA for downstream assessment. We use Adam optimizer and MSE loss to fine-tune the pre-trained UniQA. We employ a learning rate of 2e-4 for the adapter and 2e-6 for the UniQA backbone if the backbone is not frozen. We follow the typical training strategy to fine-tune each dataset, including random cropping and random horizontal flipping. Since different datasets have different MOS scales, we scale their range to [0, 1] through normalization. Considering the data scale, we train 50 epochs for LIVE, CSIQ and CLIVE datasets, 10 epochs for AVA and BIAD datasets, and 20 epochs for other datasets. During inference, we typically crop an input image into 10 image patches and take their average as the quality score of this image [2], [27]. We use the resolution of 224 × 224 for training and testing. All fine-tuning experiments are performed on an A100 GPU.

D. Main Results

Results on IQA task. Table I reports the performance of the SOTA IQA methods on seven typical IQA datasets. Our method demonstrates impressive performance across a diverse range of datasets, fully confirming the effectiveness of our method in precisely characterizing image quality. We can also observe that unfreezing the UniQA backbone for training improves performance effect, demonstrating the powerful quality characterization capability of UniQA. And fine-tuning the adapter which uses only 0.2% of the number of parameters required for unfreezing UniQA backbone, yields highly competitive performance compared to other methods. This shows the efficiency and effectiveness of our proposed adapter. Furthermore, our method demonstrates more substantial advancements on synthetic data compared to authentic data, particularly evident on the TID2013 [82] and KADID [17] with its diverse distortion types and large data sizes. We attribute this phenomenon to the enhanced commonsense knowledge and perception of visual quality acquired through extensive quality and aesthetics pre-training.

Results on IAA task. We report the experimental results on the AVA, AADB and BAID datasets in Table III, Table IV and Table V, respectively. Given that the pre-trained model acquired a unified and robust image assessment perception, it can also achieve SOTA results after fine-tuning on these three datasets. These results validate that our method can be effectively applied to both IQA and IAA domains.

E. Generalization Capability Validation

Cross dataset validation. Table VII evaluate the generalization capability of our model. We directly utilize the pre-trained UniQA and textual prompts for quality assessment. It is more challenging than other methods as the model isn't optimized on MOS labels. As observed, our method achieves the best performance on these three datasets. Notably, our method shows excellent performance on AIGC images (AGIQA-3K [83]), which are markedly different from natural images. These results demonstrate the strong generalization capability of our UniQA. Additionally, the UniQA outperforms the original CLIP significantly, proving the effectiveness of our unified pre-training.

Zero-shot image retrieval. We use different text queries to calculate the image-text similarity and rank them for zero-shot image retrieval. Fig. 4 shows the top retrieval results. We notice that "good image" prompts retrieve sharp, aesthetically pleasing images, whereas "bad image" prompts retrieve blurry, poor lighting and meaningless images. These examples provide qualitative evidence of the quality and aesthetic knowledge captured by the pre-trained model.

More downstream datasets. Our UniQA can be an effective foundation model for various downstream datasets. We evaluate on three datasets, *i.e.*, AI-generated IQA dataset AIGC-3K [83] (Table VIIIa), underwater IQA dataset UWIQA [84] (Table VIIIb) and medical IQA dataset ECIQAD [85] (Table VIIIc). We can observe that our UniQA can achieve SOTA performance on these three datasets. Since images on these three datasets come from completely different scenarios, it is very challenging to consistently achieve the leading performance on all of them. Correspondingly, these observations fully confirm that our UniQA has powerful generalization ability and can be used as an effective baseline method for various IQA scenarios.

F. Data-Efficient Learning

The pre-trained model acquires extensive image assessment knowledge, providing robust priors for downstream tasks. Consequently, our model can deliver impressive performance with limited data. To validate this, we randomly select subsets of 50, 100, and 200 samples from the training set for training and evaluate them on the same test data as full-data supervised learning. We report the median performance across 10 times

TABLE VIII: Applying UniQA to AIGC-3K [83], UWIQA [84] and ECIQAD [85] datasets. Our UniQA can be used as a foundational model for other downstream tasks and achieves SOTA performance

(a) Results on the AIGC-3K dataset

Method	SRCC	PLCC
NIQE [64]	0.524	0.567
DBCNN [66]	0.821	0.876
HyperNet [2]	0.836	0.890
CLIPIQA [52]	0.843	0.805
Q-Align [10]	0.673	0.691
MUSIQ [3]	0.826	0.866
PSCR [88]	0.850	0.906
Ours	**0.876**	**0.917**
Ours†	**0.888**	**0.922**

(b) Results on the UWIQA dataset

Method	SRCC	PLCC
FDUM [84]	0.694	0.689
UCIQE [89]	0.627	0.626
URanker [90]	0.674	0.663
UIQM [91]	0.595	0.589
UIQI [92]	0.742	0.741
CSN [92]	0.784	0.753
HPUIQA [92]	0.796	0.790
Ours	**0.847**	**0.859**
Ours†	**0.876**	**0.889**

(c) Results on the ECIQAD dataset

Method	SRCC	PLCC
BRISQUE [64]	0.436	0.459
BIQME [93]	0.770	0.768
BPRI [94]	0.152	0.181
FRIQUEE [95]	0.663	0.656
CIQA [96]	0.738	0.735
ECIQ [3]	0.839	0.842
LIQE [3]	**0.878**	-
Ours	0.873	**0.887**
Ours†	**0.918**	**0.928**

TABLE IX: SRCC results under data-efficient learning setting. * denotes using ensemble prompts

	CLIVE			KonIQ			LIVE		
Method	50	100	200	50	100	200	50	100	200
HyperIQA [2]	0.648	0.725	0.790	0.615	0.710	0.776	0.892	0.912	0.929
TReS [68]	0.670	0.751	0.799	0.713	0.719	0.791	0.901	0.927	0.957
CLIP [48]	0.664	0.721	0.733	0.736	0.770	0.782	0.896	0.923	0.941
CLIPIQA [52]	0.646	0.611	0.642	0.579	0.620	0.667	0.633	0.724	0.784
Re-IQA [26]	0.591	0.621	0.701	0.685	0.723	0.754	0.884	0.894	0.929
DEIQT [27]	0.667	0.718	0.812	0.638	0.682	0.754	0.920	0.942	0.955
GRepQ [24]	0.760	0.791	0.822	**0.812**	0.832	0.855	0.926	0.937	0.953
Ours	**0.813**	**0.836**	**0.850**	0.772	**0.842**	**0.870**	**0.962**	**0.956**	**0.974**
Ours*	**0.828**	**0.849**	**0.853**	**0.844**	**0.860**	**0.876**	**0.963**	**0.958**	**0.976**

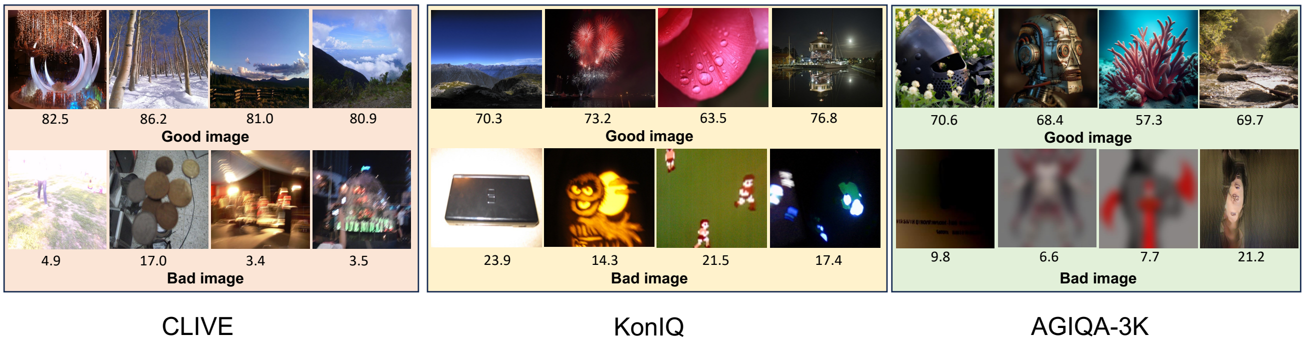


Fig. 4: The image retrieval results on three dataset with varied prompts. The number below the image is its MOS label.

in Table IX. Our method notably outperforms the second-best model GRepQ by a substantial margin, even though GRepQ is specifically designed for data-efficient learning. These results thoroughly demonstrate the potent capability of our method to learn image quality in few-label setting. Additionally, several insightful observations can be drawn from Table IX. Firstly, the prompt ensemble strategy significantly boosts model performance in data-efficient settings by fully utilizing pre-trained model knowledge. Secondly, its impact on synthetic datasets is slight, likely due to limited image variety of synthetic images, making a single prompt adequate.

G. Ablation Studies

Impact of different pre-training data. Table X explores the impact of different pre-training data. We observe that unified pre-training achieves the optimal performance on both tasks. Additionally, we derive some meaningful observations. (1) Using generated Y_{IQA} or Y_{IAA} improves the performance of both IQA and IAA tasks, proving the mutual benefit of two tasks and the effectiveness of MLLMs captioning. (2) Unifying Y_{IQA} and Y_{IAA} datasets does not lead to significant improvements. We believe this is because the MLLMs-generated text tends to have similar sentence structures [97] and representations, limiting the diversity provided for multi-modal learning. (3) Pre-training with refined human-annotated Y_{IAA}^{+} shows significant improvement on two tasks, indicating its comprehensive and effective representation for the model.

Fig. 5 illustrates the Grad-CAM visualization of different pre-training. We notice that after quality and aesthetic pre-training, the model pays more attention to blurred subjects and noisy backgrounds. This effect becomes more pronounced with unified pre-training, underscoring the advantages of such a unified approach. In addition, the unified pre-training can

TABLE X: ABLATION ON IQA (CLIVE AND KONIQ) AND IAA (AVA) DATASETS WITH SRCC METRICS

Ablation type			CLIVE	KonIQ	AVA
Ablation on different pre-training data					
Y_{IQA}	Y_{IAA}	Y^{+}_{IAA}			
×	×	×	0.865	0.907	0.748
✓	×	×	0.871	0.914	0.755
×	✓	×	0.871	0.917	0.755
✓	✓	×	0.874	0.918	0.756
×	×	✓	0.875	0.928	0.773
×	✓	✓	0.877	0.930	0.774
✓	✓	✓	**0.890**	**0.933**	**0.776**
Ablation on data purification strategy					
w/o Strategy			0.876	0.929	0.772
IR Strategy			0.879	0.931	0.774
AR Strategy			0.885	0.930	0.774
AIR Strategy			**0.890**	**0.933**	**0.776**
Ablation on different IR strategies					
Shannon Entropy			0.886	**0.934**	0.773
Sentence Length			**0.890**	0.933	**0.776**
Ablation on the proposed adapter					
Single Prompt			0.705	0.920	0.765
Antonym Prompt			0.875	0.928	0.771
Ours adapter			**0.890**	**0.933**	**0.776**
Ablation on different MLLMs					
LLaVA-v1.5-7B			0.871	0.914	0.755
LLaVA-v1.5-13B			0.872	0.914	0.757
Sphinx			0.874	0.916	0.758
QWen-VL			0.870	0.913	0.757
LLaVA-7B+QWen			0.875	0.916	0.758
Sphinx+QWen			**0.877**	**0.918**	**0.759**

focus on the areas of quality and aesthetic pre-training at the same time. This shows that unified training can learn common representations of the two tasks.

Effectiveness of data purification strategy. The second part of Table X illustrates the ablation study of the data purification strategy. It can be observed that employing either AR or IR strategy to purify data can improve the model's performance of both IQA and IAA tasks. These results validate the benefit of obtaining aesthetically relevant and informative descriptions for the model. Finally, when combining these two strategies, the best performance is achieved.

Discussion about the informativeness rank. The third part of Table X explores the effectiveness of different informativeness rank (IR) strategies. We can observe that using other informative methods (*e.g.*, Shannon Entropy) yields similar performance to our IR. This is because, in addition to IR, we also use aesthetic relevance metrics to reflect the content quality of the text. Thus, even using simple sentence length as the informativeness measure can achieve satisfactory results.

Effectiveness of the Multi-Cue Integration Adapter. The fourth part of Table X evaluates the proposed adapter. "Single Prompt" denotes using the similarity between the text "good image" and images as the score directly, while "Antonym Prompt" represents using the relative weights of texts "good image" and "bad image" to weight the predefined score. It is evident that the "Single Prompt" is considerably inferior to the "Antonym Prompt", showing the limitations of using semantic similarity as score directly. Our method integrates more cues into the "Antonym Prompt" to comprehensively assess images, thereby achieving optimal performance.

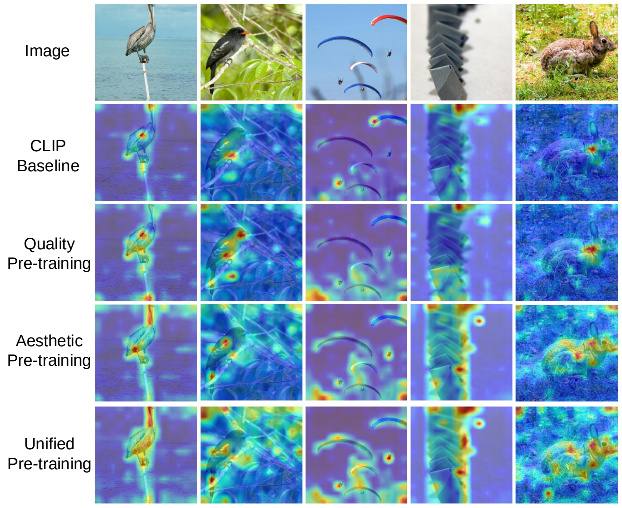


Fig. 5: Grad-CAM [98] visualization of different pre-training for prompt "blurry image". Through pre-training, the model focuses more on noisy objects and backgrounds.

Ablation on different MLLMs. The bottom part of Table X presents the ablation study of various MLLMs. We generate Y_{IQA} via different MLLMs for pre-training. It is evident that using different MLLMs exhibits similar performance, while ensembling different MLLMs can boost performance. This indicates that MLLMs are capable of generating accurate captions with our text-guided prompt, and enhancing caption diversity can further improve performance. Considering resource limitations, we use LLaVA-7B and will integrate more MLLMs in the future.

VI. CONCLUSION

This paper introduces UniQA, which leverages unified vision-language pre-training to address quality and aesthetic assessment problems concurrently. We construct a high-quality image-text dataset about quality and aesthetics via MLLMs. Through large-scale pre-training on this dataset, UniQA learns shared and effective representations of IQA and IAA tasks, enhancing the performance of two tasks significantly. In addition, we propose a Multi-Cue Integration Adapter to effectively adapt the pre-trained UniQA to downstream assessment tasks. Our method achieves state-of-the-art performance on both IQA and IAA tasks, and demonstrates powerful zero-shot and few-label image assessment capabilities.

Limitations and future work. MLLMs often generate captions with similar sentence structures and semantic expressions, restricting their ability to provide diverse and enriched representations for multimodal learning. Future work will explore other techniques to address this issue, including integrating various MLLMs for captioning and employing in-context learning methods.

References

[1] S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, "Refinednet: A weakly supervised refinement framework for single image dehazing," *IEEE Transactions on Image Processing*, vol. 30, pp. 3391–3404, 2021.

[2] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR*, 2020, pp. 3667–3676.

[3] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *ICCV*, 2021, pp. 5148–5157.

[4] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *IJCAI*, 2022, pp. 942–948.

[5] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang, "Vila: Learning image aesthetics from user comments with vision-language pretraining," in *CVPR*, 2023, pp. 10 041–10 051.

[6] F. Götz-Hahn, L.-K. Wong, and V. Hosu, "The inter-relationship between photographic aesthetics and technical quality," in *Modeling Visual Aesthetics, Emotion, and Artistic Style*. Springer, 2023, pp. 231–255.

[7] M. Jenadeleh, M. M. Masaeli, and M. E. Moghaddam, "Blind image quality assessment based on aesthetic and statistical quality-aware features," *Journal of Electronic Imaging*, vol. 26, no. 4, pp. 043 018–043 018, 2017.

[8] K. Zhang, D. Zhu, X. Min, Z. Gao, and G. Zhai, "Synergetic assessment of quality and aesthetic: Approach and comprehensive benchmark dataset," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2536–2549, 2023.

[9] X. Sheng, L. Li, P. Chen, J. Wu, L. Xu, Y. Yang, and Y. Li, "Technical quality-assisted image aesthetics quality assessment," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 50–62.

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

[11] T. Wei, H. Chen, and Q. Jiang, "Dual-stream interaction network for assessing image technique and aesthetic quality," *IEEE Transactions on Consumer Electronics*, 2024.

[12] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE TIP*, vol. 30, pp. 3474–3486, 2021.

[13] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE TIP*, vol. 15, no. 11, pp. 3440–3451, 2006.

[14] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE TIP*, vol. 25, no. 1, pp. 372–387, 2015.

[15] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.

[16] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *CVPR*, 2020, pp. 3677–3686.

[17] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[18] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *CVPR*, 2020, pp. 3575–3585.

[19] K. Ghosal, A. Rana, and A. Smolic, "Aesthetic image captioning from weakly-labelled photographs," in *ICCV Workshops*, 2019, pp. 0–0.

[20] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[22] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen *et al.*, "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models," *arXiv preprint arXiv:2311.07575*, 2023.

[23] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[24] S. Srinath, S. Mitra, S. Rao, and R. Soundararajan, "Learning generalizable perceptual representations for data-efficient no-reference image quality assessment," in *WACV*, 2024, pp. 22–31.

[25] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?" *IEEE TMM*, vol. 21, no. 5, pp. 1221–1234, 2018.

[26] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *CVPR*, 2023, pp. 5846–5855.

[27] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," *arXiv preprint arXiv:2304.04952*, 2023.

[28] K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, and W. Lin, "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2662–2672.

[29] Z. Yu, F. Guan, Y. Lu, X. Li, and Z. Chen, "Sf-iqa: Quality and similarity integration for ai generated image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6692–6701.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[31] X. Li, T. Gao, X. Zheng, R. Hu, J. Zheng, Y. Shen, K. Li, Y. Liu, P. Dai, Y. Zhang *et al.*, "Adaptive feature selection for no-reference image quality assessment using contrastive mitigating semantic noise sensitivity," *arXiv preprint arXiv:2312.06158*, 2023.

[32] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*. Springer, 2006, pp. 288–301.

[33] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, vol. 1. IEEE, 2006, pp. 419–426.

[34] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *CVPR*. IEEE, 2011, pp. 33–40.

[35] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *ICCV*, 2015, pp. 990–998.

[36] J. Hou, S. Yang, and W. Lin, "Object-level attention for aesthetic rating distribution prediction," in *ACM MM*, 2020, pp. 816–824.

[37] T. Shi, C. Chen, Z. Wu, A. Hao, and Y. Fang, "Improving image aesthetic assessment via multiple image joint learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[38] Y. Huang, L. Li, P. Chen, J. Wu, Y. Yang, Y. Li, and G. Shi, "Coarse-to-fine image aesthetics assessment with dynamic attribute selection," *IEEE Transactions on Multimedia*, 2024.

[39] S. He, A. Ming, S. Zheng, H. Zhong, and H. Ma, "Eat: An enhancer for aesthetics-oriented transformers," in *ACM MM*, 2023, pp. 1023–1032.

[40] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *CVPR*, 2021, pp. 8475–8484.

[41] J. Duan, P. Chen, L. Li, J. Wu, and G. Shi, "Semantic attribute guided image aesthetics assessment," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5.

[42] J. Hou, H. Ding, W. Lin, W. Liu, and Y. Fang, "Distilling knowledge from object classification to aesthetics assessment," *IEEE TCSVT*, vol. 32, no. 11, pp. 7386–7402, 2022.

[43] L. Li, Y. Huang, J. Wu, Y. Yang, Y. Li, Y. Guo, and G. Shi, "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE TCSVT*, 2023.

[44] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE TMM*, vol. 23, pp. 611–623, 2020.

[45] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang, "Joint image and text representation for aesthetics analysis," in *ACM MM*, 2016, pp. 262–266.

[46] X. Nie, B. Hu, X. Gao, L. Li, X. Zhang, and B. Xiao, "Bmi-net: A brain-inspired multimodal interaction network for image aesthetic assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5514–5522.

[47] Y. Huang, X. Sheng, Z. Yang, Q. Yuan, Z. Duan, P. Chen, L. Li, W. Lin, and G. Shi, "Aesexpert: Towards multi-modality foundation model for image aesthetics perception," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5911–5920.

[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[49] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," *arXiv preprint arXiv:2111.07783*, 2021.

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

[51] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[52] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, vol. 37, no. 2, 2023, pp. 2555–2563.

[53] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *CVPR*, 2023, pp. 14 071–14 081.

[54] S. Hentschel, K. Kobs, and A. Hotho, "Clip knows image aesthetics," *Frontiers in Artificial Intelligence*, vol. 5, p. 976235, 2022.

[55] X. Sheng, L. Li, P. Chen, J. Wu, W. Dong, Y. Yang, L. Xu, Y. Li, and G. Shi, "Aesclip: Multi-attribute contrastive learning for image aesthetics assessment," in *ACM MM*, 2023, pp. 1117–1126.

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

[57] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[58] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai *et al.*, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," *arXiv preprint arXiv:2309.14181*, 2023.

[59] Y. Huang, Q. Yuan, X. Sheng, Z. Yang, H. Wu, P. Chen, Y. Yang, L. Li, and W. Lin, "Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception," *arXiv preprint arXiv:2401.08276*, 2024.

[60] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *International Scholarly Research Notices*, vol. 2013, 2013.

[61] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

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

[63] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[64] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[65] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE TIP*, vol. 27, no. 1, pp. 206–219, 2017.

[66] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2018.

[67] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *CVPR*, 2020, pp. 14 143–14 152.

[68] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *WACV*, 2022, pp. 1220–1230.

[69] Z. Pan, H. Zhang, J. Lei, Y. Fang, X. Shao, N. Ling, and S. Kwong, "Dacnn: Blind image quality assessment via a distortion-aware convolutional neural network," *IEEE TCSVT*, vol. 32, no. 11, pp. 7518–7531, 2022.

[70] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.

[71] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *ECCV*. Springer, 2022, pp. 459–479.

[72] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *CVPR*, 2019, pp. 9375–9383.

[73] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE TIP*, vol. 29, pp. 3898–3910, 2020.

[74] L. Li, J. Duan, Y. Yang, L. Xu, Y. Li, and Y. Guo, "Psychology inspired model for hierarchical image aesthetic attribute prediction," in *ICME*. IEEE, 2022, pp. 1–6.

[75] L. Celona, M. Leonardi, P. Napoletano, and A. Rozza, "Composition and style attributes guided image aesthetic assessment," *IEEE TIP*, vol. 31, pp. 5009–5024, 2022.

[76] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 879–886.

[77] H. Zhu, L. Li, J. Wu, S. Zhao, G. Ding, and G. Shi, "Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1798–1811, 2020.

[78] R. Yi, H. Tian, Z. Gu, Y.-K. Lai, and P. L. Rosin, "Towards artistic image aesthetics assessment: a large-scale dataset and a new method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 388–22 397.

[79] Y. Wang, W. Cao, N. Sheng, H. Shi, C. Guo, and Y. Ke, "Tsc-net: theme-style-color guided artistic image aesthetics assessment network," in *Computer Graphics International Conference*. Springer, 2023, pp. 193–203.

[80] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*. IEEE, 2012, pp. 2408–2415.

[81] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.

[82] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database tid2013: Peculiarities and preliminary results," in *European workshop on visual information processing (EUVIP)*. IEEE, 2013, pp. 106–111.

[83] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *arXiv preprint arXiv:2306.04717*, 2023.

[84] N. Yang, Q. Zhong, K. Li, R. Cong, Y. Zhao, and S. Kwong, "A reference-free underwater image quality assessment metric in frequency domain," *Signal Processing: Image Communication*, vol. 94, p. 116218, 2021.

[85] G. Yue, D. Cheng, T. Zhou, J. Hou, W. Liu, L. Xu, T. Wang, and J. Cheng, "Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5549–5561, 2023.

[86] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*. Springer, 2016, pp. 662–679.

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

[88] J. Yuan, X. Cao, L. Cao, J. Lin, and X. Cao, "Pscr: Patches sampling-based contrastive regression for aigc image quality assessment," *arXiv preprint arXiv:2312.05897*, 2023.

[89] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[90] C. Guo, R. Wu, X. Jin, L. Han, W. Zhang, Z. Chai, and C. Li, "Underwater ranker: Learn which is better and how to be better," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 1, 2023, pp. 702–709.

[91] G. Hou, S. Zhang, T. Lu, Y. Li, Z. Pan, and B. Huang, "No-reference quality assessment for underwater images," *Computers and Electrical Engineering*, vol. 118, p. 109293, 2024.

[92] Y. Liu, K. Gu, J. Cao, S. Wang, G. Zhai, J. Dong, and S. Kwong, "Uiqi: A comprehensive quality evaluation index for underwater images," *IEEE Transactions on Multimedia*, 2023.

[93] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 4, pp. 1301–1313, 2017.

[94] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2017.

[95] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of vision*, vol. 17, no. 1, pp. 32–32, 2017.

[96] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, X. Meng, and Y.-S. Ho, "Perceptual quality assessment of cartoon images," *IEEE Transactions on Multimedia*, vol. 25, pp. 140–153, 2021.

[97] Y. Liu, K. Wang, W. Shao, P. Luo, Y. Qiao, M. Z. Shou, K. Zhang, and Y. You, "Mllms-augmented visual-language representation learning," *arXiv preprint arXiv:2311.18765*, 2023.

[98] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[99] H. Zhou, R. Yang, Y. Zhang, H. Duan, Y. Huang, R. Hu, X. Li, and Y. Zheng, "Unihead: unifying multi-perception for detection heads," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[100] H. Zhou, R. Yang, R. Hu, C. Shu, X. Tang, and X. Li, "Etdnet: Efficient transformer-based detection network for surface defect detection," *IEEE transactions on instrumentation and measurement*, vol. 72, pp. 1–14, 2023.

[101] Y. Xiao, Y. Ma, S. Li, H. Zhou, R. Liao, and X. Li, "Semanticac: semantics-assisted framework for audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[102] X. Chu, H. Zhou, Y. Zhang, Y. Zhang, R. Hu, H. Duan, Y. Huang, Y. Zheng, and R. Ji, "Attention-driven acoustic properties learning for underwater target ranging," *Pattern Recognition*, vol. 164, p. 111560, 2025.

[103] L. Tang, Z. Tian, K. Li, C. He, H. Zhou, H. Zhao, X. Li, and J. Jia, "Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 346–365.

[104] H. Zhou, R. Hu, and X. Li, "Video object segmentation with dynamic query modulation," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.