






ATTIQA: Generalizable Image Quality Feature Extractor using Attribute-aware Pretraining

Daekyu Kwon¹, Dongyoung Kim¹, Sehwan Ki², Younghyun Jo²,
Hyong-Euk Lee², and Seon Joo Kim¹

¹ Yonsei University

² Samsung Advanced Institute of Technology

Abstract. In no-reference image quality assessment (NR-IQA), the challenge of limited dataset sizes hampers the development of robust and generalizable models. Conventional methods address this issue by utilizing large datasets to extract rich representations for IQA. Also, some approaches propose vision language models (VLM) based IQA, but the domain gap between generic VLM and IQA constrains their scalability. In this work, we propose a novel pretraining framework that constructs a generalizable representation for IQA by selectively extracting quality-related knowledge from VLM and leveraging the scalability of large datasets. Specifically, we select optimal text prompts for five representative image quality attributes and use VLM to generate pseudo-labels. Numerous attribute-aware pseudo-labels can be generated with large image datasets, allowing our IQA model to learn rich representations about image quality. Our approach achieves state-of-the-art performance on multiple IQA datasets and exhibits remarkable generalization capabilities. Leveraging these strengths, we propose several applications, such as evaluating image generation models and training image enhancement models, demonstrating our model’s real-world applicability.

Keywords: Image Quality Assessment · Vision Language Model

1 Introduction

No-reference image quality assessment (NR-IQA) [9, 21, 35, 36, 39, 48] is a task of quantifying the quality of images without a pristine reference image. Recently, methods in IQA have also started incorporating deep learning [2, 14, 19, 41, 43, 44], similar to other fields in computer vision. However, effective application of deep learning in image quality assessment (IQA) faces challenges due to the limited size of existing IQA datasets [5, 8, 12, 26, 40]. Training an IQA model from scratch with a small dataset encounters difficulties in learning rich representations for image quality. This often results in degraded performance and poor generalization, thereby restricting the practical applicability of IQA in real-world scenarios.

To address the generalization issue derived from limited dataset size, IQA approaches have been developed to leverage rich representations from large datasets [30] (Fig. 1(a)). In [1, 36, 38, 44], transfer learning strategy was utilized

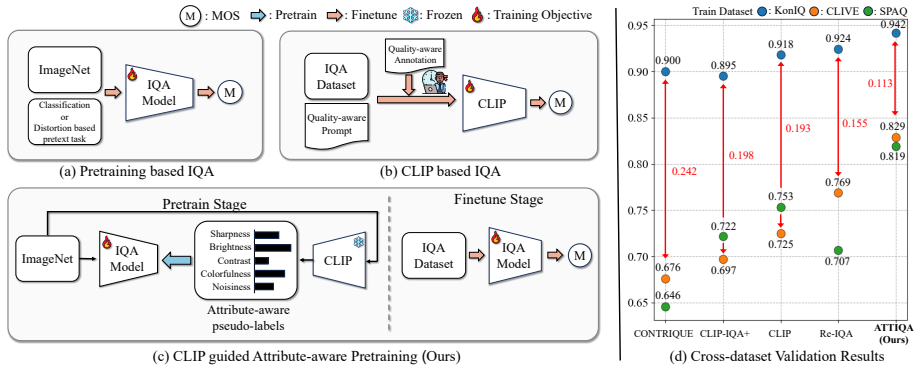


Fig. 1: An illustration of a training strategy for IQA across previous works and ours. (a) Classic IQA models use ImageNet-pretrained models or suggest image quality-related pretraining. (b) CLIP-based IQA directly utilizes CLIP or adapts it for IQA using additional quality annotations, which requires human labor. (c) Our method incorporates the rich representation of large datasets and leverages CLIP’s IQA capability. We pre-train IQA model with attribute-aware pseudo-labels derived from CLIP and finetune it to the target IQA dataset. (d) Cross dataset validation results, obtained by testing on the KonIQ dataset after training on various datasets. ATTIQA achieves state-of-the-art results and exhibits superior generalization capability on unseen datasets, showing less performance decline on cross-dataset setup compared to other methods.

by pretraining a model on ImageNet [30]. Several studies [19, 20, 32, 44, 46] have proposed IQA-specific pretext tasks, founded on the premise that distortions in images directly impact their quality. These lines of research emphasize the importance of pretraining tasks in IQA, demonstrating the benefits of the scalability of large datasets. However, research on how to efficiently extract quality-related representations from large datasets is still in progress.

Recently, Vision Language Models (VLM), exemplified by CLIP [27], have emerged as a robust backbone in computer vision, highlighting their generalization capabilities. Building on these strengths, exploiting VLM for IQA has also been explored (Fig. 1(b)). CLIP-IQA [37] proposed zero-shot IQA using the quality-aware prompt, showing the applicability of CLIP for IQA. While CLIP-based IQA demonstrates good generalization capability without fine-tuning, it has been noted that CLIP alone is not suitable for precise IQA tasks, as it is trained on generic image-text pairs. To address this issue, several studies attempt to adapt CLIP to the IQA domain using text prompts related to image quality. Although these approaches effectively enhance CLIP’s representation for IQA, these strategies are constrained by the necessity of supplementary image-text pairs for direct CLIP training, which requires additional human labor.

In this work, we introduce a novel pretraining framework for IQA, named “**ATTIQA**”, **AT**Tribute-aware **IQA**, which exhibits enhanced generalization capabilities by effectively incorporating CLIP’s extensive knowledge and the scalability of large unlabeled datasets. While previous works [37, 45] have observed

that CLIP inherently contains robust representations relevant to IQA, the representation of CLIP also consists of a wide range of semantic contexts, hindering the precise assessment of image qualities. To this end, we propose a pretraining scheme that distills only quality-aware knowledge from CLIP with unlabeled large dataset. Specifically, we generate pseudo-labels for given unlabeled images utilizing CLIP’s zero-shot inference with quality-aware prompts and use them for training a target encoder (Fig. 1(c)). Such a pretraining scheme can effectively transfer CLIP’s quality-related knowledge, along with the scalability benefits of unlabeled large datasets, into the target encoder. This results in the construction of robust representations that contain only helpful information for IQA.

To generate quality-aware pseudo-labels, we propose to incorporate prompts based on five key attributes, which have been proven to be crucial for assessing image quality [5, 13, 34]. Specifically, we propose a five image attribute based pretraining strategy beyond Mean Opinion Score(MOS). Instead of using generic prompts such as “a good/bad photo”, as used in CLIP-IQA [37], we select prompts representing each key attribute through Large Language Model(LLM) and our carefully designed proxy tasks. They are taken by CLIP as inputs to generate pseudo-labels, facilitating a network to learn from five representation spaces for each specific image attribute (Fig. 1(c)). Taking advantage of using a large-scale dataset combined with the novel attribute-aware CLIP guidance, our pretraining framework significantly enhances the learning of rich representations closely associated with image attributes and quality. We demonstrate the effectiveness of our method through extensive experiments, achieving state-of-the-art performance on multiple IQA datasets, as well as on an aesthetic quality dataset.

The ability to generalize beyond the training dataset is crucial for IQA, particularly when considering its further applications. We observe that ATTIQA exhibits superior performance when the test dataset is unseen (Fig. 1(d)) or the training dataset is limited, which is more applicable to real-world scenarios. Building on these strengths, we propose a couple of applications where a generalizable IQA method can be employed. We show that our method can be used to evaluate the outputs of a generative model [29] and as a reward function for reinforcement learning-based image enhancement [33].

2 Related Work

Classical Image Quality Assessment. Since image quality is highly regarded as essential in diverse vision applications, numerous image quality assessment studies have been explored. Traditional NR-IQA utilizes a feature-based machine learning approach to quantify image quality, leading to a primary focus on extracting meaningful features. Therefore, these works introduced hand-crafted feature based IQA, which is derived from natural scene statistics [7], spatial domain [21, 22] or frequency domain [31].

Deep learning based IQA. With the success of deep learning, various deep learning-based NR-IQA methods have been introduced. Early works tried to train neural networks by directly predicting mean opinion score (MOS) [14]

or the distribution of MOS [36]. Some works attempted to incorporate meta-learning [48] or hypernetwork [35]. However, the limited size of IQA datasets restricts deep learning-based approaches, making it hard to extract rich representations solely relying on the IQA dataset. To address this problem, recent IQA methods [36, 38, 44] commonly adopted ImageNet [30] classification-based backbone as their initial state, which already possesses rich representations. However, there is another problem that this representation is not fully suitable to IQA, as their pretraining task mainly focuses on semantic information but not image quality.

Pretrain based IQA. Beyond applying deep learning strategies to IQA, some approaches have focused on generating quality-aware representation using large datasets without the need for ground truth. Liu *et al.* [19] employed the Siamese network to rank images according to their quality, generating images of varying quality levels by applying different scales of distortion to a single image. Synthetic distortion-aware representation was introduced in [44], which attempts to classify the type or the amount of distortions applied to images. Recently, with the success of SSL, some works [20, 32, 46] suggested a contrastive learning framework refined for IQA. Unlike typical contrastive learning, they viewed patches from the same image, and each patch is differently distorted as different-quality samples. By treating these samples as negative pairs in the training process, they efficiently constructed a quality-aware representation space.

Vision Language Model based IQA. VLM [18, 28] is a foundation model that learns correspondence between image and text to understand the relationships between visual contents and language. Specifically, CLIP [27] is trained with 400 million image-text pairs, and thus, it shows generalization capability for various computer vision tasks. Taking advantage of this ability, IQA methods that utilize CLIP have also been introduced. CLIP-IQA [37] directly assessed image quality by measuring the image’s correspondence with quality-aware text prompts. They also suggested an enhanced version named CLIP-IQA+ that optimizes text prompts to adapt to the given target dataset. Despite this successful application, since the CLIP is trained with unrefined caption data focusing on semantic information, there is still room for improvement in refining CLIP’s representation space towards an image quality-aware representation space. In response to this property, some works have tried to adapt the representation space of CLIP with additional datasets. Ke *et al.* [16] fine-tuned the CoCa [42] with the aesthetic captioning dataset to make aesthetic-aware VLM. By injecting aesthetic-related vision-language correspondence into VLM, they showed improved performance of their representation in quality-related downstream tasks. Zhang *et al.* [45] suggested a multi-task learning approach to adapt the vision backbone of VLM for a unified IQA dataset. They trained the vision backbone by optimizing cosine similarities of multi-modal embeddings with task-aware prompts.

As demonstrated by the above approaches, works refining the representation of VLM to aware image quality are currently underway. While these methodologies showed improvements in incorporating quality-aware information into VLM’s representation space, the necessity of additional datasets to fine-tune re-

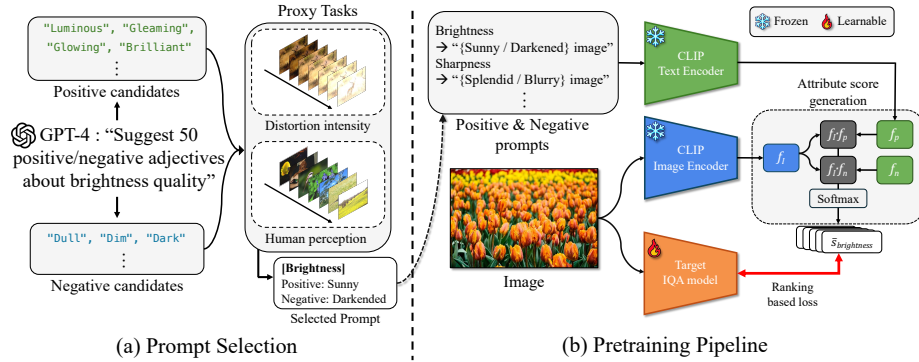


Fig. 2: (a) The overall process of our prompt selection strategy for each image attribute (e.g. brightness). Given attribute, we create prompt candidates using GPT-4 and then find the optimal prompt by utilizing proxy tasks related to the attribute. (b) ATTIQA’s proposed pretraining pipeline. We generate attribute scores using CLIP with an antonym strategy and then train our target IQA model using ranking-based loss with generated scores.

mains a limitation. To mitigate this issue, our method does not *fine-tune* the CLIP model itself, but rather *extract* quality-related information from CLIP and use them to train our IQA model.

3 Method

Fig. 2 illustrates our framework, which consists of two primary components: prompt selection and pretraining pipeline using pseudo-labels from CLIP [27]. During the prompt selection phase, we create a list of candidate prompts using a large language model (LLM). We then identify the most suitable prompts for generating image attribute scores by evaluating the score generation ability of candidates through proxy tasks. In the pretrain stage, we generate image attribute scores as pseudo-labels for pretraining using CLIP and the chosen prompts. Our IQA model is pretrained on this pseudo-labeled data and fine-tuned using a dedicated IQA dataset.

3.1 Image Attributes

Our method aims to utilize image attribute-based scores as supervision to reduce the ambiguity of the IQA task, which is typically represented solely by MOS. This approach yields a more precise and well-defined representation of image quality by offering specific quality criteria beyond the generic and ambiguous mere notions of ‘good’ and ‘bad’.

In the line of IQA research that aims to incorporate quality relevant information beyond the MOS, five key attributes – *sharpness*, *contrast*, *brightness*,

colorfulness, and *noisiness*— are widely employed and have proven beneficial for the IQA task [5, 13, 34]. Especially, SPAQ [5] substantiated through a user study that these five image attributes correlate well with perceived image quality. These observations indicate that these attributes are helpful factors in understanding the image quality. Therefore, we choose these five attributes as the target objective for our IQA model. Note that the following attribute score generation and prompt selection are conducted separately for each image attribute.

3.2 Attribute Score Generation

During attribute score generation, we generate five attribute scores for given images using CLIP’s zero-shot inference. Given image x and attribute-aware prompt t , CLIP encodes the image and prompt into shared multi-modal feature space. We then compute the relatedness score s between x and t using cosine similarity as follows:

$$s(x, t) = \frac{E_I(x) \cdot E_T(t)^T}{\|E_I(x)\| \cdot \|E_T(t)\|}, \quad (1)$$

where E_I and E_T represent CLIP’s image and text encoder, respectively.

Our pseudo-label generation employs an antonym strategy [37], which computes scores by integrating scores of positive and negative prompts with the softmax function. For example, we can use a prompt pair {“*Dark image*”, “*Bright image*”} as a negative-positive pair to calculate the brightness attribute score. Then, our attribute score is computed by the following equation:

$$\bar{s}_{attribute}(x) = \frac{e^{s(x, t_{pos})}}{e^{s(x, t_{pos})} + e^{s(x, t_{neg})}}, \quad (2)$$

where t_{pos} and t_{neg} represent positive and negative prompts for the corresponding image attribute, respectively.

3.3 Prompt Selection

Previous work involving CLIP [47] has shown that the choice of prompts is critical in determining performance. To address this, we introduce a prompt selection strategy aimed at identifying the most effective prompts for generating image attribute scores. Drawing inspiration from techniques used in the NLP field [6], we develop a selection based efficient approach for identifying the optimal prompts.

As depicted in the left side of Figure 2(a), we begin by generating prompt candidates using Large Language Model (LLM), specifically GPT-4 [25]. To streamline the search process, we adopt a standard template for these prompts in the format “[*adjective*] image”, focusing specifically on a variation of adjectives. Prompt candidates are constructed using GPT-4, with an ask query to elicit adjectives pertinent to specific image attributes. Since we use antonym pair for each attribute, we generate 50 positive and 50 negative adjectives, resulting in 2500 positive-negative prompt candidates for each attribute. Subsequently, we identify

the most suitable prompt pair from 2500 candidates by assessing their capability to generate accurate attribute scores. To find the optimal prompt pair, we present two proxy tasks. The optimal prompt is determined as the prompt that produces an attribute score that best aligns with the goals of both tasks. To measure an image attribute appropriately, the proxy tasks are designed under two hypotheses: 1) For a fixed image, when a distortion corresponding to the image attribute is applied to it, the attribute score predicted by the prompt should increase or decrease accordingly. 2) For different images, the attribute scores generated by the prompt pairs should match well with the degree of human perception of each attribute.

In the first proxy task, we apply varying levels of distortion to a fixed image and identify the optimal pair of prompts that yield an attribute score aligning most accurately with the applied level of distortion. For the second task, we employ the SPAQ dataset, which comprises diverse scenes and offers human-annotated attribute scores for the same five attributes we adopt. We aim to identify the prompt pair that generates the attribute scores whose order closely aligns with the order of the provided scores in the dataset. We calculate the sum of SROCC scores for both tasks and select the highest performing prompt pair, and the result is shown in Table 1. These selected pairs are then utilized to generate each attribute score in the following pretraining pipeline.

Table 1: Results of the prompt selection. These prompts are chosen by our prompt selection strategy.

Attribute	Positive prompt	Negative prompt
Sharpness	" <i>Splendid image</i> "	" <i>Blurry image</i> "
Contrast	" <i>Distinct image</i> "	" <i>Vague image</i> "
Brightness	" <i>Sunny image</i> "	" <i>Darkened image</i> "
Colorfulness	" <i>Vibrant image</i> "	" <i>Colorless image</i> "
Noisiness	" <i>Peerless image</i> "	" <i>Graimy image</i> "

3.4 Attribute Aware Pretraining Pipeline

After selecting prompts, we train the target IQA model to construct attribute-aware space with ranking-based loss using a pseudo-label derived from CLIP, as illustrated in Fig. 2(b).

Our method aims to create five unique representation spaces for each specific image attribute. Accordingly, our IQA model comprises a shared encoder backbone and five attribute heads for each image attribute. Each attribute head consists of two-layer multi-layer perceptrons (MLPs) that output an attribute score. Then, our training objective for the pretraining pipeline can be formulated by minimizing the discrepancy between five attribute score predictions from the IQA model and the corresponding image attribute scores generated from CLIP.

Our pretraining pipeline can be directly implemented using regression-based loss such as MSE or L1 loss. However, directly using the attribute score with regression-based loss hinders handling uncertainties. Since scoring image attributes as scalars in a zero-shot manner is inherently challenging, training by predicting these scores may impede the construction of robust representations.

To address this problem, we use a relative ranking-based loss instead of a numerical norm-based loss to minimize the dependence on CLIP’s numerical results, which are subject to uncertainty. To implement this loss in our framework, we utilize margin-ranking loss that optimizes the relative ranking of the two samples given. We first define the indicator function F , which specifies the superiority of image attribute based on its score \bar{s} for given images:

$$F_a(x_1, x_2) = \begin{cases} 0, & \bar{s}_a(x_1) > \bar{s}_a(x_2) \\ 1, & \bar{s}_a(x_1) \leq \bar{s}_a(x_2), \end{cases} \quad (3)$$

where a denotes an element of image attributes set $A = \{\textit{sharpness}, \textit{contrast}, \textit{brightness}, \textit{colorfulness}, \textit{noisiness}\}$, and x_1 and x_2 denote the sample images.

Then, we train our target IQA model based on margin-ranking loss with the indicator function F_a . We compute our loss by summation of margin-ranking loss independently for each attribute a using attribute score prediction from respective attribute head E_a :

$$L = \sum_a \max(0, m - (E_a(x_1) - E_a(x_2))) \cdot F_a(x_1, x_2), \quad (4)$$

where m denotes the margin hyper-parameter.

3.5 Fine-tuning

To predict MOS with our IQA model, we have to fine-tune it on the target IQA dataset. However, the architecture of our IQA model during the pretraining stage is not designed to output a single score but instead predicts five attribute scores. Therefore, we adapt the architecture of our model to fit the IQA task to generate a single MOS as output. At a fine-tuning stage, we extract features from each attribute head, excluding the final layer, and these features are then concatenated and fed into regression MLPs that predict the MOS.

4 Experiments

4.1 Datasets

We conduct experiments with ATTIQA on the 4 “in-the-wild” NR-IQA Datasets, CLIVE [8], KonIQ-10k [12], SPAQ [5], and FLIVE [40] and 1 image aesthetic dataset, AVA [23]. While AVA [23] focuses on image aesthetics, we utilize this dataset since its user study setting is the same as “IQA in the wild”.

For FLIVE [40] and AVA [23], we follow the official dataset split. For the rest, we randomly partition the dataset and allocate 80% to the train set and 20% to the test set. Following previous works [46], we conduct the same experiment ten times with different random splits and report the median value as a result to compensate for the bias arising from random splits.

Table 2: Fine-tuning performance comparison of ATTIQA and existing NR-IQA methods for 4 IQA “in-the-wild” dataset and 1 IAA dataset. “†” denotes that this measurement is achieved from [46]. “-” denotes that measurement is not possible due to the absence of an official code and result. Other measurements are based on the official reports or reproduced by the official code. We highlight the best performance in bold and underline the second-best performance for each dataset.

Methods	CLIVE		KonIQ		SPAQ		FLIVE		AVA	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
DBCNN [†] [44]	0.844	0.862	0.878	0.887	0.906	0.907	0.542	0.626	0.554	0.583
HyperIQA [†] [35]	0.855	0.871	0.908	0.921	0.916	0.919	0.535	0.623	0.668	0.668
CONTRIQUE [20]	0.824	0.848	0.900	0.915	0.910	0.915	0.598	0.674	0.674	0.678
MUSIQ [15]	-	-	0.916	0.928	0.917	0.921	0.646	0.739	0.726	0.738
TReS [9]	0.846	0.877	0.915	0.928	0.915	0.919	0.554	0.625	0.658	0.663
Re-IQA [32]	0.823	0.865	0.924	0.935	0.915	0.919	0.574	0.674	0.714	0.716
QPT [46]	<u>0.895</u>	0.914	0.927	0.941	<u>0.925</u>	<u>0.928</u>	0.610	0.677	-	-
CLIP [27]	0.847	0.881	0.918	0.932	0.918	0.922	0.563	0.628	0.746	0.745
CLIP-IQA+ [20]	0.805	0.832	0.895	0.909	0.864	0.866	0.575	0.593	0.692	0.732
LIQE [45]	0.865	0.866	0.898	0.913	-	-	-	-	-	-
ATTIQA (Distortion intensity)	0.891	0.910	<u>0.929</u>	<u>0.943</u>	0.922	0.926	0.625	0.729	0.754	0.750
ATTIQA (Human perception)	0.890	<u>0.915</u>	0.942	0.952	<u>0.925</u>	0.930	<u>0.635</u>	<u>0.740</u>	<u>0.756</u>	<u>0.759</u>
ATTIQA (Joint strategy)	0.898	0.916	0.942	0.952	0.926	0.930	0.632	0.742	0.761	0.761

4.2 Experimental Setup

For a fair comparison, we utilize ResNet-50 [10] as our backbone, widely used in NR-IQA. For CLIP, we adopt the ViT-B/16 model [4]. We experimentally set the value of the margin parameter m at the loss function to 0.1.

At the pretraining stage, we use the ImageNet [30] widely used for pretext tasks. At the fine-tuning stage, we followed the setting from [46]. We resized the image’s shorter edge to 340 and randomly cropped the image at a resolution of 320×320 . We fine-tuned our network to 100 epochs on each target dataset. At the evaluation stage, we take five crops at a resolution of 320×320 from each corner and center as test samples, and the average of the results is used for the predicted MOS. For performance evaluation, we calculated Pearson’s Linear Correlation Coefficient (PLCC) and Spearman’s Rank-Order Correlation Coefficient (SROCC), widely adopted evaluation metrics in IQA research.

4.3 Main Result

In this section, we report the quantitative performance of ATTIQA and compare it with existing NR-IQA models. Utilizing our CLIP-guided attribute-aware pre-trained model, we conduct fine-tuning on five IQA datasets to predict the MOS. As shown in the Table 2, ATTIQA shows notable performance improvements compared to CLIP-based methods on four IQA “in the wild” datasets and one image aesthetic dataset AVA. Our method demonstrates state-of-the-art performance in most evaluation settings, with the second-best SROCC performance on the FLIVE dataset. It is important to note that MUSIQ is an exceptional work

Table 3: Cross dataset validation performance comparison of ATTIQA and existing NR-IQA methods.

Train DB	CLIVE			KonIQ			SPAQ			FLIVE		
Test DB	KonIQ	SPAQ	FLIVE	CLIVE	SPAQ	FLIVE	CLIVE	KonIQ	FLIVE	CLIVE	KonIQ	SPAQ
CONTRIQUE	0.676	0.842	0.346	0.731	0.789	0.410	0.549	0.646	0.338	0.706	0.709	0.734
Re-IQA	0.769	0.852	0.424	0.791	0.862	0.461	0.732	0.707	0.497	0.720	0.676	0.793
CLIP	0.725	0.850	0.405	0.799	0.837	0.507	0.773	<u>0.753</u>	0.496	0.727	0.717	0.834
CLIP-IQA+	0.697	0.836	0.437	0.803	0.832	0.516	0.784	0.722	0.470	0.620	0.631	0.661
LIQE	<u>0.819</u>	<u>0.877</u>	<u>0.497</u>	<u>0.824</u>	<u>0.868</u>	0.551	-	-	-	-	-	-
ATTIQA	0.829	0.887	0.511	0.856	0.879	<u>0.540</u>	0.824	0.819	0.548	0.756	0.762	0.867

that utilized the complete FLIVE dataset comprising patches and full images, unlike the other methods that do not use patch data. Notably, ATTIQA shows a significant performance gap on the KonIQ-10k and AVA datasets.

In the last three rows of Table 2, we report the performance of three different versions of ATTIQA. We carry out experiments with various types of prompts, including cases where we apply the two proxy tasks described in Sec 3.3—Distortion Intensity and Human Perception—separately, as well as a scenario where we combine both tasks in our prompt selection strategy (Joint Strategy). We observe that prompts based on *human perception* work effectively, and the *joint strategy* that involves both proxy tasks shows the most superior performance. It indicates that a prompt selection strategy that considers using both low-level information *distortion* and high-level *human perception* enhances the robustness of our model across various datasets.

4.4 Generalization Capability

Cross-dataset Validation. To verify ATTIQA’s generalization ability, we conduct experiments about cross-dataset validation. This experiment evaluates the IQA model’s ability to learn generalizable features by training it on the specified dataset and testing it on the unseen dataset. To consider various scenarios, we conduct extensive experiments across four datasets: CLIVE, KonIQ, SPAQ, and FLIVE. Every experimental setup is the same as the main experiment, and due to the various ranges of the MOS for each dataset, we use only SROCC as an evaluation criterion. As shown in Table 3, ATTIQA exhibits superior generalization capability to baselines, achieving the best performance in most scenarios. Interestingly, LIQE achieves comparable results to ATTIQA, demonstrating that strategies adapting CLIP possess strong generalization capabilities. However, we highlight that ATTIQA outperforms LIQE in most scenarios and that LIQE cannot be extended to datasets where additional annotations are not provided.

Data-Efficient Setup. Moreover, we conducted experiments in a data-efficient setup to demonstrate that ATTIQA can generalize in environments where only a small amount of data is available. Instead of the conventional 8:2 Train-Test split, we performed training using only 10% or 20% of the data. Since we utilize a small amount of data, we also use only SROCC as an evaluation criterion.

As shown in Table 4, ATTIQA outperforms other pretrain-based methods in environments with limited datasets. This performance gap validates that our pretraining strategy is more robust than other baselines.

Table 4: Comparison of ATTIQA and NR- **Table 5:** Cosine similarity between IQA methods which focuses on representation features from pretrained and fine-tuning under data efficient setup. tuned encoder.

Methods	CLIVE		KonIQ		SPAQ		Fine-tune DB	CLIVE	KonIQ	SPAQ
	10%	20%	10%	20%	10%	20%				
CONTRIQUE	0.687	0.740	0.832	0.835	0.883	0.885	CONTRIQUE	0.536	0.566	0.613
Re-IQA	0.632	0.683	0.853	0.888	0.893	0.902	Re-IQA	0.158	0.181	0.243
CLIP	0.650	0.728	0.846	0.863	0.882	0.884	CLIP	0.195	0.203	0.309
ATTIQA	0.820	0.838	0.903	0.919	0.909	0.917	ATTIQA	0.890	0.811	0.945

Feature Analysis. In this section, we analyze why ATTIQA exhibits superior generalization capability compared to other pretrain-based methods. We hypothesize that pretext tasks providing more generalizable representations would offer robust features that do not overfit specific datasets. To examine these properties, we extract features from each backbone before and after fine-tuning the IQA dataset and compare them by measuring cosine similarity. As shown in Table 5, ATTIQA’s features are slightly adjusted, whereas other methods’ features are modified significantly. This result suggests that ATTIQA’s pretrained representation inherently possesses superior robustness and provides a more effective initialization point for IQA than other methods, leading to enhanced performance and generalization capability.

For real-world applications, a model’s generalization ability is far more critical than its performance on specific benchmark datasets. Our method’s superior generalization capability ensures robust baseline performance on unseen images, highlighting its practicability when considering the purpose of the IQA tasks. Building on this strength, we will demonstrate the application of ATTIQA in real-world scenarios in Sec 5.

4.5 Ablation Studies

Linear probing. We conduct linear probing experiments to demonstrate the robustness of our attribute-aware pretrained feature space, training only a single regression MLPs on the target dataset with a frozen ATTIQA backbone. In this experiment, we compared ATTIQA to previous works that suggest pretext tasks for IQA. For Re-IQA, we report the three types of results based on the encoder configuration: using only the quality or content encoder, and both encoders. As shown in Table 6, ATTIQA shows a significant performance gap compared to other methods in CLIVE. In other datasets, ATTIQA also demonstrates comparable results to Re-IQA, which uses both features of a separate quality and content encoder, while our method only utilizes a single shared encoder for five

attributes. We note that CLIP shows the worst performance, validating our motivation that CLIP’s original representations are unsuitable for precise IQA.

Table 6: Linear probing performance comparison of ATTIQA and NR-IQA methods which focuses on representation learning.

Methods	CLIVE		KonIQ		SPAQ	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
CONTRIQUE	0.845	0.857	0.894	0.906	0.916	0.919
Re-IQA (quality)	0.806	0.824	0.861	0.885	0.900	0.910
Re-IQA (content)	0.808	0.844	0.896	0.912	0.902	0.908
Re-IQA (both)	0.840	0.854	0.914	0.923	0.918	0.925
CLIP	0.803	0.829	0.883	0.895	0.895	0.896
ATTIQA	0.870	0.891	<u>0.903</u>	<u>0.918</u>	0.918	<u>0.922</u>

Table 7: Ablation study results about our prompt based strategy and loss function.

Prompt type	CLIVE		KonIQ		SPAQ	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
ATTIQA	0.898	0.916	0.942	0.952	0.926	0.930
single-prompt	0.880	0.909	0.928	0.939	0.916	0.920
Worst prompt	0.869	0.889	0.930	0.943	0.920	0.925
Median prompt	0.872	0.893	0.931	0.944	0.921	0.925
with L_2	0.875	0.904	0.933	0.945	0.923	0.928

Attribute based Approach. To demonstrate the effectiveness of our image attribute based approach, we carry out an experiment by replacing the target objective from five image attributes with a single overall image quality. In this experiment, we train the IQA model with a single pseudo-label using a prompt pair that describes image quality: $\{“Good image”, “Bad image”\}$. Comparing the first and the second tab of Table 7, we can observe that our representation space decomposing image quality into five attributes outperforms the single-prompt based representation space, justifying our approach for model design.

Prompt Selection Strategy. To justify our prompt selection strategy, we experiment with other prompts selected by different strategies. For comparison, we adopt prompts that achieve the median and lowest scores in the proxy task. As shown in the third tab of Table 7, the results of our strategy align with the performance at the evaluation. This correlation validates the efficacy of our prompt selection strategy.

Ranking-based loss. To verify the efficacy of our relative ranking-based loss approach, we conduct an additional ablation study by replacing the margin-ranking loss with L2 loss at the pretraining stage. As depicted in the last row of Table 7, the use of L2 loss exhibits a performance degradation compared to adopting margin-ranking loss. Interestingly, we can observe a notable performance decline in the CLIVE dataset, which has the smallest dataset size within this ablation study. This result supports the use of relative loss instead of numerical loss, enhancing our pretraining pipeline’s robustness.

5 Applications

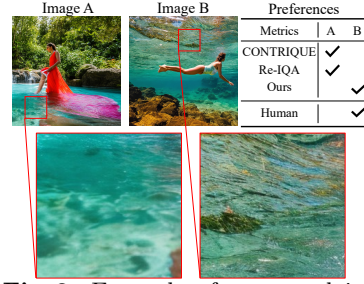
To better demonstrate our ATTIQA’s generalization capability, we introduce two types of applications in this section: (1) metrics for the generative model and (2) image enhancement guided by our IQA score. For each application, we employ models trained on the KonIQ dataset, which shows the best generalization capability at Sec 4.4.

Table 8: Comparisons of accuracy between human preferences and IQA model’s result.

Method	CONTRIQUE	Re-IQA	CLIP-IQA+	ATTIQA
Accuracy (%)	61.5	55.0	57.5	71.0

Table 9: Performance comparisons among IQA model in an AI-Generated Contents Dataset(AGIQA-3k)

Method	CONTRIQUE	Re-IQA	CLIP-IQA+	ATTIQA
SROCC	0.643	0.807	0.835	0.854
PLCC	0.795	0.876	0.885	0.911

**Fig. 3:** Example of generated images. The images are generated by the same prompt. ATTIQA hits human preference while others do not.

5.1 Metrics for Generative Model

Recently, as the diffusion models [11, 29] have shown success in the text-to-image generation [24, 28] task. One of their primary focus is generating high quality images from a given text prompt. In this regard, we attempt to employ ATTIQA as a metric for generative models.

To validate ATTIQA’s effectiveness as a metric, we create a benchmark dataset that involves the pairwise comparison of two images generated from the same text prompt. Here, we generate 200 pairs of images using the Stable Diffusion [29] and collect human preference by conducting a user study. When collecting the user preferences, we only present the generated images without the prompt to make participants focus on visual quality. The user study was carried out with 60 participants through Amazon Mechanical Turk (AMT). We then investigate the correlation between IQA models and the human participants.

As shown in Table 8 and Figure 3, our method mostly aligns with human preference compared to other IQA methods. Our ATTIQA can capture this detailed visual quality difference while others do not. Please refer to the supplementary for the user study details and more visual results. We will make the benchmark used in this application publicly available for further IQA research.

Moreover, we carry out an additional experiment using an AGIQA-3k dataset [17], which consists of images generated by various generative models. As shown in Table 9, ATTIQA outperforms other methods, exhibiting a significant performance gap. These results highlight the improved generalization capability of our method when extended to AI-generated content. They indicate the potential for expanding the use of ATTIQA as a metric to evaluate generative models.

5.2 Image Enhancement

The image signal processing pipeline (ISP) converts an input raw image into a color image. It is essential to carefully tune the parameters of the ISP to obtain visually pleasing images. In this section, we apply ATTIQA’s MOS prediction as

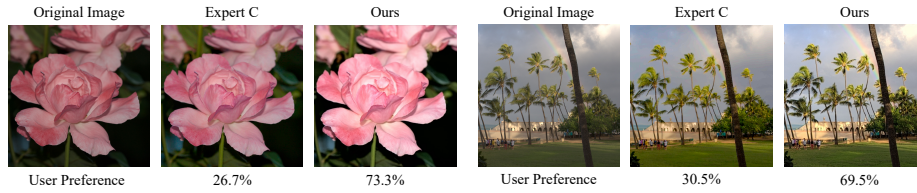


Fig. 4: Qualitative comparisons between our enhancement method and retouching of Expert C. Our results give more liveliness and vibrancy, aligned more closely with human preference.

a reward for reinforcement learning to find optimal parameters for the ISP [33]. After the training, we convert raw images into color images in the MIT-Adobe-5k dataset [3], which consists of 5,000 raw images and color images retouched by five experts (A/B/C/D/E). Then, we conduct a user study comparing our result against the retouched one by expert C, which is typically used as the ground truth in most previous image enhancement research. The study was executed with 60 participants through AMT, involving a comparison of 200 image pairs. For details on the implementation, please refer to the supplementary materials.

As shown in Fig. 4, our pipeline retouches images to make them more colorful and vivid compared to both retouching by expert C and the default settings. Furthermore, according to our user study, ATTIQA receives higher preferences from subjects, demonstrating a 58% win rate compared to Expert C. We also report additional qualitative comparisons to supplementary material.

6 Discussion and Conclusion

We propose ATTIQA, a pretraining framework for IQA that develops an attribute-aware representation space with CLIP guidance. Since our IQA model effectively incorporates CLIP’s vast knowledge and scalability of large datasets, it shows state-of-the-art performance on IQA datasets and superior generalization capability on cross-dataset validation. Leveraging these advantages, we successfully demonstrate a couple of real-world applications where IQA can be utilized.

Limitation and Future Work. While our approach focuses on five attributes commonly employed in the IQA domain, we expect that other properties relevant to image quality exist (e.g., Composition and Focus). Consequently, future work will involve exploring extended representation spaces for IQA. Given that our method proposes a pretraining framework not limited to the specified attributes, our work holds the potential for expansion to encompass additional properties.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (Artificial Intelligence Graduate School Program, Yonsei University, under Grant 2020-0-01361).

References

1. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing* **12**, 355–362 (2018) [1](#)
2. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing (TIP)* **27**(1), 206–219 (2017) [1](#)
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2011) [14](#)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)* (2021) [9](#)
5. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3677–3686 (2020) [1](#), [3](#), [6](#), [8](#)
6. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) *Proceedings of the Association for Computational Linguistics (ACL)*. pp. 3816–3830 (2021) [6](#)
7. Gao, X., Gao, F., Tao, D., Li, X.: Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. *IEEE Transactions on neural networks and learning systems* (2013) [3](#)
8. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing (TIP)* **25**(1), 372–387 (2015) [1](#), [8](#)
9. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 1220–1230 (2022) [1](#), [9](#)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016) [9](#)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 6840–6851 (2020) [13](#)
12. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing (TIP)* **29**, 4041–4056 (2020) [1](#), [8](#)
13. Huang, Y., Li, L., Yang, Y., Li, Y., Guo, Y.: Explainable and generalizable blind image quality assessment via semantic attribute reasoning. *IEEE Transactions on Multimedia (TMM)* (2022) [3](#), [6](#)
14. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1733–1740 (2014) [1](#), [3](#)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 5148–5157 (2021) [9](#)

16. Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., Yang, F.: Vila: Learning image aesthetics from user comments with vision-language pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10041–10051 (2023) [4](#)
17. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: Agiqa-3k: An open database for ai-generated image quality assessment (2023) [13](#)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning (ICML). pp. 12888–12900 (2022) [4](#)
19. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Rankiqa: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1040–1049 (2017) [1](#), [2](#), [4](#)
20. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing (TIP)* **31**, 4149–4161 (2022) [2](#), [4](#), [9](#)
21. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing (TIP)* **21**(12), 4695–4708 (2012) [1](#), [3](#)
22. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters* **20**(3), 209–212 (2012) [3](#)
23. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2408–2415 (2012) [8](#)
24. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International conference on machine learning (ICML) (2021) [13](#)
25. OpenAI: Gpt-4 technical report (2023) [6](#)
26. Ponomarenko, N.N., Jin, L., Ieremeiev, O., Lukin, V.V., Egiazarian, K.O., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., Kuo, C.C.J.: Image database tid2013: Peculiarities, results and perspectives. *Signal Processing, Image Communication* **30**, 57–77 (2015) [1](#)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (ICML) (2021) [2](#), [4](#), [5](#), [9](#)
28. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning (ICML). pp. 8821–8831 (2021) [4](#), [13](#)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) [3](#), [13](#)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* pp. 211 – 252 (2014) [1](#), [2](#), [4](#), [9](#)
31. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing (TIP)* **21**(8), 3339–3352 (2012) [3](#)

32. Saha, A., Mishra, S., Bovik, A.C.: Re-iqa: Unsupervised learning for image quality assessment in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5846–5855 (2023) 2, 4, 9
33. Shin, U., Lee, K., Kweon, I.S.: Drl-isp: Multi-objective camera isp with deep reinforcement learning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7044–7051 (2022) 3, 14
34. Su, S., Hosu, V., Lin, H., Zhang, Y., Saupe, D.: Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In: British Machine Vision Conference (BMVC) (2021) 3, 6
35. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3667–3676 (2020) 1, 4, 9
36. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing (TIP) 27(8), 3998–4011 (2018) 1, 4
37. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 37, pp. 2555–2563 (2023) 2, 3, 4, 6
38. Yang, X., Li, F., Liu, H.: Ttl-iqa: Transitive transfer learning based no-reference image quality assessment. IEEE Transactions on Multimedia (TMM) 23, 4326–4340 (2021). <https://doi.org/10.1109/TMM.2020.3040529> 1, 4
39. Ye, P., Kumar, J., Kang, L., Doermann, D.: Unsupervised feature learning framework for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1098–1105 (2012) 1
40. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3575–3585 (2020) 1, 8
41. You, J., Korhonen, J.: Transformer for image quality assessment. In: IEEE International Conference on Image Processing (ICIP). pp. 1389–1393 (2021) 1
42. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models (2022) 4
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018) 1
44. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology (TCVST) 30(1), 36–47 (2018) 1, 2, 4, 9
45. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14071–14081 (2023) 2, 4, 9
46. Zhao, K., Yuan, K., Sun, M., Li, M., Wen, X.: Quality-aware pre-trained models for blind image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22302–22313 (2023) 2, 4, 8, 9
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) 130(9), 2337–2348 (2022) 6

48. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: Metaiqa: Deep meta-learning for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14143–14152 (2020) [1](#), [4](#)