Enhancing Descriptive Image Quality Assessment With a Large-Scale Multi-Modal Dataset

Zhiyuan You, Jinjin Gu, Xin Cai, Zheyuan Li, Kaiwen Zhu, Chao Dong, and Tianfan Xue

***Abstract*—With the rapid advancement of Vision Language Models (VLMs), VLM-based Image Quality Assessment (IQA) seeks to describe image quality linguistically to align with human expression and capture the multifaceted nature of IQA tasks. However, current methods are still far from practical usage. First, prior works focus narrowly on specific sub-tasks or settings, which do not align with diverse real-world applications. Second, their performance is sub-optimal due to limitations in dataset coverage, scale, and quality. To overcome these challenges, we introduce the Enhanced Depicted image Quality Assessment model (EDQA). Our method includes a multi-functional IQA task paradigm that encompasses both assessment and comparison tasks, brief and detailed responses, full-reference and non-reference scenarios. We introduce a ground-truth-informed dataset construction approach to enhance data quality, and scale up the dataset to 495K under the brief-detail joint framework. Consequently, we construct a comprehensive, large-scale, and high-quality dataset, named EDQA-495K. We also retain image resolution during training to better handle resolution-related quality issues, and estimate a confidence score that is helpful to filter out low-quality responses. Experimental results demonstrate that EDQA significantly outperforms traditional score-based methods, prior VLM-based IQA models, and proprietary GPT-4V in distortion identification, instant rating, and reasoning tasks. Our advantages are further confirmed by real-world applications including assessing the web-downloaded images and ranking model-processed images. Codes, datasets, and model weights have been released in https://depictqa.github.io/**

***Index Terms*—Image quality assessment, vision language models.**

Received 29 April 2025; revised 26 October 2025; accepted 24 November 2025. Date of publication 9 December 2025; date of current version 12 December 2025. This work was supported in part by RGC Early Career Scheme (ECS) under Grant 24209224, in part by the National Natural Science Foundation of China under Grant 62276251, and in part by the Joint Laboratory of CAS-HK. The associate editor coordinating the review of this article and approving it for publication was Prof. Kede Ma. *(Corresponding authors: Tianfan Xue; Chao Dong.)*

Zhiyuan You is with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China, and also with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: zhiyuanyou@link.cuhk.edu.hk).

Jinjin Gu is with INSAIT, Sofia University "St. Kliment Ohridski", 1784 Sofia, Bulgaria (e-mail: jinjin.gu@insait.ai).

Xin Cai is with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China (e-mail: caixin@link.cuhk.edu.hk).

Zheyuan Li is with the University of Macau, Macau, China, and also with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: zheyuanli884886@gmail.com).

Kaiwen Zhu is with Shanghai Jiao Tong University, Shanghai 200240, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: sqzhukaiwen@sjtu.edu.cn).

Chao Dong is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen University of Advanced Technology, Shenzhen 518055, China (e-mail: chao.dong@siat.ac.cn).

Tianfan Xue is with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China, also with CPII under InnoHK, Hong Kong, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: tfxue@ie.cuhk.edu.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3639998, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3639998

I. Introduction

IMAGE Quality Assessment (IQA) aims to measure and compare the quality of images, expecting to align with human perception. With the emergence of Vision Language Models (VLMs) [1], [2], [3], VLM-based IQA begins to attract more research interest [4], [5], [6], [7], [8]. These methods leverage VLMs to describe image quality using language, recognizing that language better mirrors human expression, and captures the multifaceted nature of IQA tasks [8]. However, existing VLM-based IQA methods still fall short especially in aspects of *functionality* and *performance*.

Functionality. There are various application scenarios of IQA, but existing VLM-based IQA models only support a few of them. For example, one scenario involves assessing a single image downloaded from the web, while another requires comparing multiple images handled by different algorithms. Also, image restoration needs to assess an image against a reference, while image generation requests non-reference assessments. Therefore, a superior IQA model should be multi-functional to cater to such diverse scenarios. However, existing methods limit to a specific subset of these tasks, such as single-image assessment [5], multi-image comparison [6], or full-reference setting [8], *etc.* Hence, the limitations in functionality hinder the wide applications of prior methods.

Performance. Many IQA methods perform well on some specific datasets but may generalize poorly to other images with different contents or distortions. For instance, Co-Instruct [6] performs well on TID2013 dataset [9] (85.0%), but drops significantly to 50.7% when testing on BAPPS dataset [10]. A more comprehensive comparison on our newly created benchmark is given in Fig. 1, where it shows that previous works [5], [6] under-perform even within their defined tasks and settings. One potential cause for this is the limited scope of their training datasets. For example, the added distortion category in Q-Instruct [5] is limited; Co-Instruct [6] directly utilizes GPT-4V [2], which is not accurate in IQA tasks, to generate data; and the dataset scale in DepictQA [8] remains small. Besides, these methods are constrained in their usage by resizing images to a fixed resolution [5], [6], while the image resolution is critical for quality. Therefore, the dataset's coverage, quality, and scale together with the training techniques limit the performance of previous methods.

To address these challenges, we propose a multi-functional IQA model to handle various image quality assessment tasks. We categorize these tasks into two types, as shown in Fig. 2.

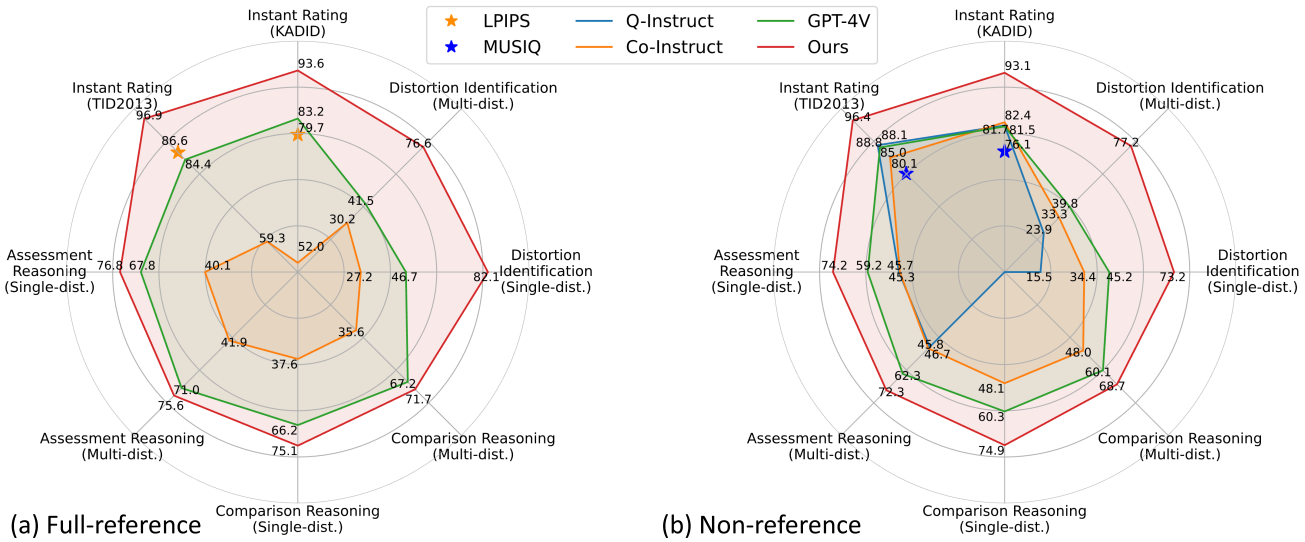


Fig. 1. **Performance comparison**. Our model surpasses previous works including Q-Instruct [5], Co-Instruct [6], and the proprietary GPT-4V [2] across a broad range of tasks in both full-reference and non-reference settings. Traditional score-based IQA methods like LPIPS [10] and MUSIQ [11] have no language abilities, and thus can only be used in *instant rating* task. Q-Instruct is only tested on single-image input tasks.

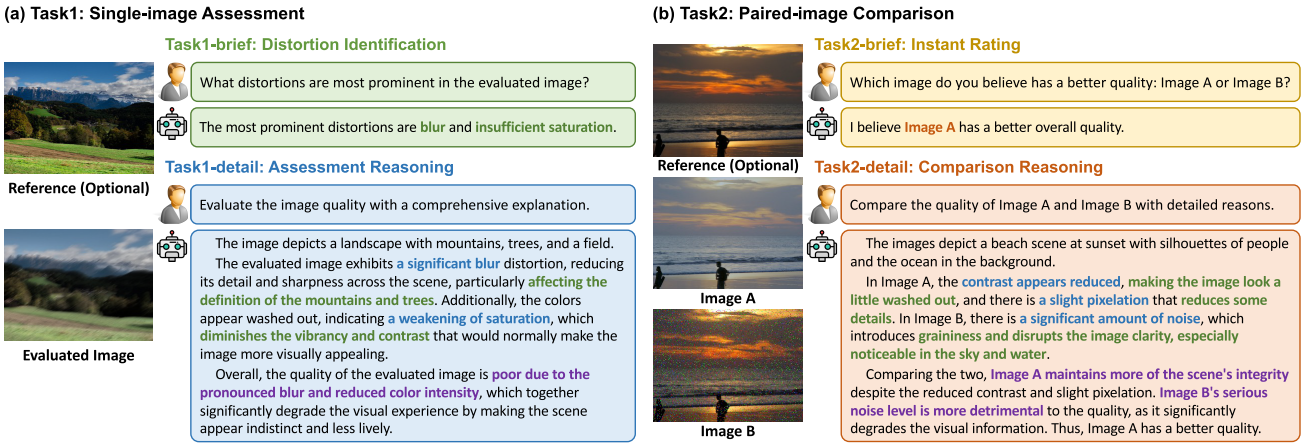


Fig. 2. **Illustration of our task paradigm and qualitative results**. Our EDQA focuses on two main tasks including *single-image assessment* and *paired-image comparison* in both *full-reference* and *non-reference* settings. Each task contains a *brief* sub-task focusing on the fundamental IQA ability, and a *detailed* sub-task fostering the reasoning capacities. More qualitative results are provided in *Supp. Mat.*

(a) *Single-image assessment* aims to evaluate the quality of a single image by identifying distortions (*e.g.*, "blur" in Fig. 2a top). It can also analyze the distortions' impacts on contents (*e.g.*, blur "affecting the definition of mountains and trees" in Fig. 2a bottom).(b) *Paired-image comparison* focuses on comparing the quality of two distorted images based on the clarity, colorfulness, and sharpness of presented contents. For example, in Fig. 2b, despite reduced contrast, "Image A maintains more scene integrity", as "Image B's serious noise level is more detrimental". We omit multi-image comparison since it is an easy extension of a pairwise one [12]. Each type includes basic *brief* sub-tasks for fundamental assessments and *detailed* sub-tasks to enhance reasoning abilities. Moreover, the model supports both *full-reference* and non-reference settings, making it adaptable to diverse scenarios.

Under the multi-functional task paradigm, we construct a new large-scale dataset, EDQA-495K, for comprehensive and accurate training and evaluation. First, for diverse distortion, we design and implement 35 types of distortions, each with 5 levels. Second, to enhance the label quality, we inform GPT-4V of the low-level ground truths (*e.g.*, distortions) to leverage its strong high-level perception and language abilities, while avoiding its sub-optimal IQA capabilities. Third, to increase the dataset scale, we scale up the data amount to 495K under the brief-detail combined framework [8]. Also, our dataset is suitable for both full-reference and non-reference settings.

With the collected EDQA-495K dataset, we then train a VLM model, named **E**nhanced **D**epicted image **Q**uality **A**ssessment model (EDQA). During training, the original image resolution is retained, leading to a better quality perception regarding resolution. Furthermore, we estimate the confidence of responses from key tokens, providing vital auxiliary information, especially for filtering low-quality responses.

The performance of EDQA is evaluated in Fig. 1 and Sec. V. In brief tasks, our model surpasses general VLMs, IQA-specific VLMs, and score-based IQA methods by a large

margin. For example, we achieve 95.9% in non-reference comparison on TID2013 dataset, remarkably surpassing Co-Instruct (85.0%) and GPT-4V (88.1%). In detailed tasks, our model also excels, *e.g.*, recording 74.9% in non-reference comparison reasoning, compared to 48.1% for Co-Instruct and 60.3% for GPT-4V. At last, experiments on real-world applications including assessing web-downloaded images and comparing model-restored images further demonstrate our superiority. We hope that our multi-functional model could serve as a stepping stone towards a unified VLM-based IQA model. Although not yet fully realized, our method showcases the potential of VLM-based IQA models.

II. Related Works

A. Score-Based IQA Methods

Traditional IQA methods rely on scores to assess image quality and can be divided into *full-reference* and *non-reference* methods.(a) Full-reference methods compute a similarity score between a distorted image and a high-quality reference. Early works rely on human-designed metrics such as image information [13], structural similarity [14], phase congruency with gradient magnitude [15], *etc.* The rapid advancement of deep learning has also inspired learning-based IQA methods that measure image quality through data-driven training. Pioneered by PieAPP [16] and LPIPS [10], data-driven approaches [17], [18], [19], [20], [21], [22], [23] have spurred innovations in IQA, exhibiting high consistency with human judgments.(b) Non-reference methods directly regress a quality score without a reference image. Initially, human-designed natural image statistics are adopted [24], [25], [26], [27], [28], [29], [30]. Subsequently, deep-learning-based methods [31], [32], [33], [34], [35], [36], [37] replace hand-crafted statistics by learning quality priors from extensive data. Recent works focus on enhancing performance by introducing multi-scale features [11], CLIP pre-training [38], multi-dimension attention [39], continual learning [40], multitask learning [41], and so on. However, as discussed in [8], score-based IQA methods limit themselves in complex analyses and multi-aspect weighing of IQA, since the information provided by a single score is far from sufficient.

Vision Language Models (VLMs) incorporate visual modality into Large Language Models (LLMs) [42], [43], [44], aiming to leverage their emergent ability to achieve general visual ability. These VLMs [1], [2], [45], [46], [47], [48], [49], [50], [51], [52] have demonstrated a general visual ability and can tackle a variety of multi-modality tasks, including image captioning [53], [54], [55], visual question answering [56], [57], [58], document understanding [59], [60], [61], *etc.* Although proficient in these high-level visual perception tasks, we demonstrate in Sec. V that general-purpose VLMs still struggle with IQA tasks.

VLM-based IQA methods aim to achieve better alignment with human perception leveraging the power of VLMs [7]. Q-Bench [4] establishes a comprehensive benchmark for evaluating general-purpose VLMs in low-level perception tasks. Reference [62] evaluates various VLMs on the widely-adopted two-alternative forced choice (2AFC) task. Q-Instruct [5] enhances the low-level perception ability of VLMs by introducing a large-scale dataset. Q-Align [63] and DeQA-Score [64] employ discrete text-defined levels for more accurate quality score regression. Co-Instruct [6] concentrates on the quality comparison among multiple images. DepictQA [8] performs quality description, quality comparison, and comparison reasoning in the full-reference setting.

Nonetheless, as highlighted in Sec. I, most of these methods focus on specific aspects of IQA tasks, diverging from the original intents of VLMs' universality and practical usage requirements.

III. Task Paradigm and Dataset Construction

A. Task Paradigm

As highlighted in the introduction, there are various application scenarios for IQA models. First, the evaluation objective can be either single-image assessment or paired-image comparison. The former is useful to rate a web-downloaded image, while the latter suits comparing images processed by two different algorithms. Second, the reference setting may be full-reference or non-reference. For example, image restoration requires assessments based on References, while image generation needs non-reference evaluations. Third, the response could be either brief or detailed. Brief responses suit well-targeted tasks (*e.g.*, comparison without reasons), while detailed responses enhance interpretability and human interaction. To cater to such diverse scenarios, a practical IQA method should be multi-functional. Therefore, we aim to establish such a multi-functional task paradigm for VLM-based IQA research. As shown in Fig. 2, we focus on two tasks, each containing both brief and detailed sub-tasks, and supporting both full-reference and non-reference settings.

- *Task1: single-image assessment*.(a) Brief sub-task: *distortion identification*. Given a distorted image, the model should identify the most obvious distortions.(b) Detailed sub-task: *assessment reasoning*. In addition to identifying distortions, the model should also describe how these distortions affect the perception of image contents and the overall image quality.
- *Task2: paired-image comparison*.(a) Brief sub-task: *instant rating*. Given two distorted images, the model should find the image with better quality.(b) Detailed sub-task: *comparison reasoning*. Building upon the comparison results, the model should first compare the content loss caused by distortions in the two images, then weigh different aspects to draw inferences, and finally justify its comparison results. Note that we omit the multi-image (> 2) comparison since it can be achieved easily as the extension of paired case [12].

Compared with previous works, our design unifies various tasks, response types, and reference settings into a multi-functional paradigm. In contrast, Q-Instruct [5] focuses on non-reference single-image assessment, Co-Instruct [6] targets comparison among multiple images in the non-reference setting, and DepictQA [8] primarily addresses the full-reference setting. Although one can achieve unified IQA by combining these task-specific IQA models, it is impractical due to the significant increase in network parameters, considering that current VLMs are already quite large.

B. Distortion Library

Existing IQA datasets (*e.g.*, BAPPS [10], PieAPP [16]) usually introduce distortions (*e.g.*, noise, blur) into high-quality

TABLE I
OVERVIEW OF OUR DISTORTION LIBRARY WITH 12 SUPER-CATEGORIES AND 35 SUB-CATEGORIES IN TOTAL

Super-category	Blur	Noise	Compression	Brighten	Darken	Contrast Strengthen	Contrast Weaken	Saturate Strengthen	Saturate Weaken	Over-sharpen	Pixelate	Quantize
# Sub-category	6	6	2	4	4	2	2	2	2	1	1	3

reference images to create distorted images for evaluation. However, these datasets do not publicly release the distortion information of each image, and their distortions only cover limited scenes. Therefore, we aim to develop a comprehensive large-scale distortion library.

1) Distortion Generation: Our distortion system comprises 12 super-categories in total, with each super-category consisting of multiple sub-categories. For instance, the "blur" category encompasses "Gaussian blur", "motion blur", "lens blur", *etc.* In total, there are 35 sub-categories. For each sub-category, there are 5 severity levels: "slight", "moderate", "obvious", "serious", and "catastrophic". A summary is illustrated in Tab. I. Considering the need to assess high-quality images as well, we retain the original image without any distortions in 5% proportion. See details in *Supp. Mat.*

2) Multi-Distortion Setups: In practical usage, multiple distortions may occur simultaneously on the same image. While a simple way to simulate them is to add multiple distortions recursively, real-world scenarios are more complex. First, one distortion may weaken another, such as "brighten" weakens "darken", "blur" weakens "over-sharpen". Second, certain distortions exhibit similar visual results, such as "pixelate" looks similar to "blur", making it challenging to identify both if they are applied simultaneously. We also observe that humans can identify at most two distortions when three or more are applied, as illustrated in *Supp. Mat.* Hence, we limit the distortion number to two and manually review all possible combinations to exclude contradictory or similar ones. See details of multi-distortion setting in *Supp. Mat.*

C. Dataset Construction

High-quality and large-scale datasets are crucial for training VLMs. Following [1], [49], training VLMs requires {images, question, response} triplets, where "images" are the ones to be evaluated, "question" describes the task, and "response" is the ground truth answer. In this section, we detail the construction of our dataset from the selection of images and the collection of questions and responses.

D. Image Collection

Typical IQA datasets involve two types of images: high-quality reference images and distorted images to be evaluated. Generating distorted images is easy given our comprehensive distortion library introduced in Sec. III-B. Existing studies often collect a large number of distorted images from a small number of references [9], [12]. However, the semantic richness of images is also crucial for VLM training. Therefore, we primarily source reference images from the KADIS-700K dataset [79], which offers 140K pristine reference images from diverse natural and daily scenes. We also leverage other IQA datasets for their convenience to generate responses (details are below).

E. Question Collection

Humans often express similar questions using different sentences, necessitating model robustness to various questions. For each task, we initially prompt GPT-4 [43] to generate 50 candidate questions. Then, we manually eliminate ambiguous and repetitive ones and correct inaccurate ones, creating a set of 20 questions. We randomly sample these questions during training and testing to form the data pair.

F. Response Collection

We employ two response types as shown in Fig. 3. The first comprises brief templated responses that are easy to produce, where we emphasize the *quantity* to bring robust fundamental skills. The second consists of detailed responses, where we emphasize the *quality* to enhance the advanced reasoning abilities. Existing methods to collect detailed responses mainly rely on human annotation [5], [8] and GPT-4V generation [6]. However, human annotation can be biased and vary in quality when annotators are untrained or tired [8]. Also, GPT-4V is not fully reliable since its IQA performance is still unsatisfactory as evidenced in Sec. V.

We rethink the key aspects of our desired responses and GPT-4V's corresponding abilities, introducing *GT-informed generation* by prompting the Ground Truth (GT) details to enhance GPT-4V's generation. Specifically, a high-quality detailed response should contain image contents, key distortions, the impacts of distortions on contents, and conclusions. While GPT-4V excels at identifying contents and analyzing impacts, it struggles with distortion identification and quality comparison, as shown in Sec. V. To compensate for that, we directly provide it with explicit GT information. The response generation for each task is detailed subsequently.

Task1-brief: distortion identification. As shown in Fig. 3a, we first establish a response pool containing 20 templates with unspecified distortions. Next, we add distortions into the reference to create its distorted counterpart and populate a sampled template with the specific distortions as the response. For streamlined evaluation, we sample half of the questions and append the short answer prompt: "Answer the question using a single word or phrase."Correspondingly, the response will be a single phrase, like "noise", specifying the distortions.

Task1-detail: assessment reasoning. Given the reference image, we initially introduce distortions to corrupt the reference. Then, GPT-4V is input with both two images and the distortion information, and requested to assess the quality of the distorted image, as illustrated in Fig. 3b. We instruct GPT-4V to respond from three dimensions: contents, distortions along with their impacts on contents, and overall quality. Prior works [5], [6] mainly focus on low-level properties, while we consider how these distortions influence the display of high-level contents.

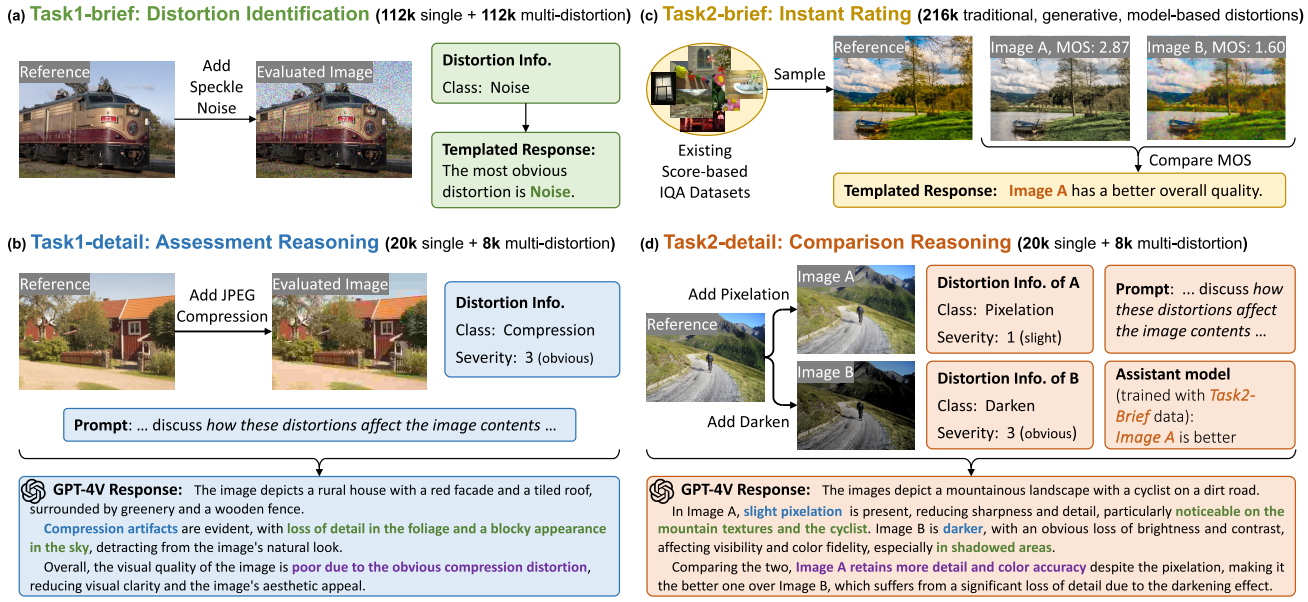


Fig. 3. **Construction of EDQA-495K dataset**. For *distortion identification*, templated responses are generated using distortion information. In *instant rating*, we sample images from existing datasets and compare the Mean Opinion Score (MOS) to determine the better image for templated response creation. For *assessment reasoning* and *comparison reasoning* tasks, we provide GPT-4V with evaluated images and Ground Truth (GT) details (*i.e.*, distortion information, comparison results from an assistant model) to facilitate detailed and accurate response generation, called *GT-informed generation*. This additional information is critical as GPT-4V cannot produce it accurately.

TABLE II
STATISTICS OF OUR EDQA-495K DATASET

(a) Number of samples. All tasks except *instant rating* are displayed in the single-distortion / multi-distortion format.

	Task1-brief *Distortion Identification*	Task1-detail *Assessment Reasoning*	Task2-brief *Instant Rating*	Task2-detail *Comparison Reasoning*
Train	112,000 / 112,000	19,829 / 7,981	215,676	19,809 / 7,958
Test	28,000 / 28,000	200 / 100	41,120	200 / 100

(b) Response length reported as word count / string length.

	Distortion Identification	Assessment Reasoning	Instant Rating	Comparison Reasoning
Single-dist.	10.36 / 69.81	64.37 / 430.23	9.30 / 52.02	93.20 / 604.97
Multi-dist.	12.84 / 88.67	87.31 / 588.44		114.04 / 740.68

Task2-brief: instant rating. We begin by sampling a reference image and its two distorted versions from existing IQA datasets, and then compare the Mean Opinion Score (MOS) to determine the better one, as shown in Fig. 3c. Similar to *distortion identification*, we assemble a response pool of 20 templates to convert the comparison results into

responses, and append the short answer prompt for the convenience of evaluation. We select three IQA datasets for training, including BAPPS [10], KADID-10K [80], and PIPAL [12], to cover a diverse range of reference images.

Task2-detail: comparison reasoning. As depicted in Fig. 3d, given a high-quality image, we randomly apply distortions to produce two distorted images. We first train an assistant model using the large-scale *instant rating* data to predict the comparison results. Note that GPT-4V does not perform well on the quality comparison task, as shown in our experiments in Tab. IV and Tab. IX, and thus we train our own comparison model. Then, similar to *assessment reasoning*, we inform GPT-4V of the three images, distortion information, and comparison results to generate detailed responses.

G. Setup of Non-Reference Setting

Our dataset accommodates both full-reference and non-reference settings. However, even for humans, identifying subtle distortions (*e.g.*, minor brightness adjustments) without a reference is challenging. Thus, in the non-reference setting, we selectively remove samples with "slight" severity on some specific distortions, including "brighten", "darken", "contrast weaken", "contrast strengthen", "saturate weaken", "saturate strengthen", "quantize", and "over-sharpen".

H. Dataset Analysis

Dataset statistics are given in Tab. IIa. Our training set contains 439,676 brief samples and 55,577 detailed samples. For *instant rating*, the training set includes BAPPS, KADID, and PIPAL, while the validation set consists of BAPPS, KADID, PIPAL, TID2013 [9], LIVE-MD [81], and MDID2013 [82]. To ensure there is no intersection between training and validation sets for those overlapped datasets, the original splits are kept. For detailed tasks, all samples in the validation set have been carefully checked by humans.

We analyze six statistical features to examine the image distributions. First, we use CLIP [83] ViT to extract semantic features and apply K-Means clustering with $k = 20$. As shown in Fig. 4a, the samples are relatively balanced across clusters. Second, the *brightness* (Fig. 4b) and *contrast* (Fig. 4c) approximately follow Gaussian distributions, consistent with natural image statistics [84]. Third, as in Fig. 4de, *colorfulness* and *edge density* exhibit right-skewed heavy-tailed distributions, which are characteristics of natural scenes [85], [86]. Finally,

TABLE III

DISTORTION IDENTIFICATION RESULTS UNDER BOTH SINGLE-DISTORTION AND MULTI-DISTORTION CASES. THE ACCURACY METRIC IS REPORTED IN THE FULL-REFERENCE / NON-REFERENCE SETTINGS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. OUR EDQA GREATLY OUTPERFORMS ALL BASELINES AND MAINTAINS ITS HIGH ACCURACY IN OUT-OF-DISTRIBUTION (OOD) SETTING

	General VLM						IQA-specific VLM					
	mPLUG-Owl2	LLaVA-1.6	LLaVA-OV-1.5	InternVL2.5	Qwen2.5-VL	GPT-4V	Q-Instruct	Co-Instruct	LIQE	Q-Insight	EDQA	EDQA-OOD
Single-dist.	10.1 / 11.6	14.0 / 15.3	23.9 / 30.4	18.7 / 20.3	33.0 / 31.8	46.7 / 45.2	-/ 15.5	27.2 / 34.4	-/ 33.1	48.4 / 49.9	**97.7 / 94.1**	82.1 / 73.2
Multi-dist.	10.8 / 10.7	12.0 / 12.1	32.6 / 38.3	35.0 / 36.8	36.5 / 38.3	41.5 / 39.8	-/ 23.9	30.2 / 33.3	-/ 31.4	43.5 / 49.6	**91.3 / 89.3**	76.6 / 77.2

TABLE IV

INSTANT RATING RESULTS ON MULTIPLE BENCHMARKS IN THE FULL-REFERENCE / NON-REFERENCE SETTING WITH THE ACCURACY METRIC. THE BEST RESULTS ARE SHOWN IN BOLD. Q-INSTRUCT IS TESTED BY INPUTTING SINGLE IMAGES TO CALCULATE QUALITY SCORES, AND THEN COMPARE THE SCORES TO RATE. EDQA SURPASSES ALL BASELINES BY A LARGE MARGIN

Methods		BAPPS^{test}	KADID^{test}	PIPAL^{test}	TID2013	LIVE-MD	MDID2013	Mean
Full-reference Score-based IQA	PSNR	68.9 /	78.7 /	80.9 /	85.0 /	89.7 /	78.0 /	80.2 /
	SSIM	69.7 /	77.1 /	82.6 /	78.7 /	88.1 /	76.8 /	78.8 /
	LPIPS	79.4 /	79.7 /	84.2 /	86.6 /	91.3 /	85.4 /	84.4 /
	DISTS	79.7 /	85.8 /	84.6 /	87.0 /	**93.1** /	88.5 /	86.5 /
Non-reference Score-based IQA	NIQE	/ 49.9	/ 66.9	/ 59.7	/ 65.0	/ 86.9	/ 82.2	/ 68.4
	ClipIQA	/ 59.7	/ 75.8	/ 72.6	/ 85.8	/ 65.8	/ 47.0	/ 67.8
	MUSIQ	/ 59.2	/ 76.1	/ 77.8	/ 80.1	/ 87.2	/ 81.1	/ 76.9
	MANIQA	/ 54.9	/ 68.4	/ 79.2	/ 77.3	/ 75.4	/ 63.5	/ 69.8
General VLM	mPLUG-Owl2	50.1 / 50.1	50.6 / 50.8	49.6 / 49.6	48.6 / 48.5	49.9 / 50.1	50.6 / 50.5	49.9 / 49.9
	LLaVA-1.6	54.1 / 56.2	50.4 / 51.9	52.0 / 52.6	54.2 / 57.0	54.4 / 56.5	54.3 / 53.1	53.2 / 54.6
	LLaVA-OV-1.5	50.0 / 49.7	50.1 / 51.4	50.2 / 49.6	51.9 / 49.8	51.3 / 47.8	50.3 / 52.3	50.6 / 50.1
	InternVL2.5	50.5 / 62.8	50.2 / 75.5	50.3 / 67.0	52.4 / 83.9	52.0 / 73.0	49.4 / 73.2	50.8 / 72.6
	Qwen2.5-VL	49.6 / 55.3	55.8 / 76.2	56.7 / 67.9	61.9 / 83.3	54.5 / 64.3	50.8 / 55.6	54.9 / 67.1
	GPT-4V	70.3 / 63.2	83.2 / 81.5	78.5 / 78.2	84.4 / 88.1	79.6 / 72.7	70.6 / 67.6	77.8 / 75.2
IQA-specific VLM	Q-Instruct	- / 41.6	- / 81.7	- / 74.6	- / 88.8	- / 73.1	- / 48.5	- / 68.1
	Co-Instruct	49.8 / 50.7	52.0 / 82.4	50.6 / 72.5	59.3 / 85.0	50.0 / 70.3	50.0 / 58.0	52.0 / 69.8
	Q-Insight	49.9 / 50.8	51.1 / 85.0	63.4 / 68.6	54.2 / 90.8	48.9 / 50.7	55.8 / 58.0	53.9 / 67.3
	EDQA (Ours)	**84.7 / 82.4**	**93.6 / 93.1**	**90.5 / 90.0**	**96.9 / 96.4**	92.1 / **91.8**	**90.0 / 89.6**	**91.3 / 90.6**

TABLE V

ASSESSMENT REASONING AND COMPARISON REASONING RESULTS WITH GPT-4 SCORE METRIC IN THE FULL-REFERENCE / NON-REFERENCE SETTING. THE BEST RESULTS ARE MARKED IN BOLD.

Methods	Assessment Reasoning		Comparison Reasoning	
	Single-dist.	Multi-dist.	Single-dist.	Multi-dist.
GPT-4V	67.8 / 59.2	71.0 / 62.3	66.2 / 60.3	67.2 / 60.1
Q-Instruct	- / 45.7	- / 45.8	-	-
Co-Instruct	40.1 / 45.3	41.9 / 46.7	37.6 / 48.1	35.6 / 48.0
EDQA (Ours)	**76.8 / 74.2**	**75.6 / 72.3**	**75.1 / 74.9**	**71.7 / 68.7**

texture variance (Fig. 4f) follows a highly right-skewed γ distribution, aligning with prior findings [87]. These results confirm that our dataset maintains natural image statistics in semantic and low-level dimensions.

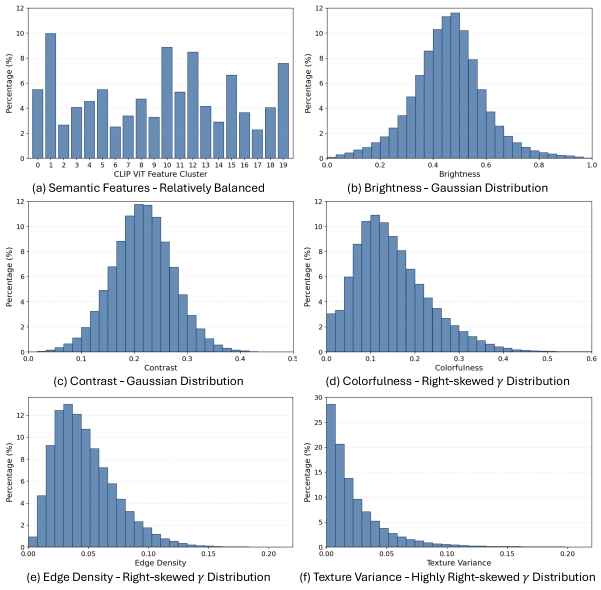


Fig. 4. Statistics of all images in our dataset about (a) semantic features, (b) brightness, (c) contrast, (d) colorfulness, (e) edge density, and (f) texture variance.

IV. MODEL DESIGN

A. Model Architecture

EDQA primarily adopts the architecture from LLaVA-1.6 [88] and mPLUG-Owl2 [3], structured as follows. Specifically, the input images and the question texts are first tokenized, then fused, finally processed by the Large Language Model (LLM) for response generation. (1) Tokenizing input images and question texts. We use a frozen CLIP pre-trained ViT-L/14 [83] as the image encoder to convert the input images into visual tokens. The question texts are tokenized into textual tokens using the SentencePiece tokenizer [89]. To bridge the different embedding spaces of visual and textual tokens, we

TABLE VI

ASSESSMENT REASONING AND COMPARISON REASONING RESULTS WITH CLASSICAL METRICS (BLEU AND ROUGE-L) IN THE FULL-REFERENCE / NON-REFERENCE SETTING. THE BEST RESULTS ARE BOLD.

Methods	Assessment Reasoning				Comparison Reasoning			
	Single-distortion		Multi-distortion		Single-distortion		Multi-distortion	
	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L
GPT-4V	0.020/0.010	0.248/0.224	0.023/0.015	0.246/0.223	0.029/0.025	0.261/0.243	0.030/0.024	0.251/0.238
Q-Instruct	- / 0.003	- / 0.210	- / 0.002	- / 0.198	-	-	-	-
Co-Instruct	0.008/0.005	0.201/0.204	0.002/0.003	0.209/0.203	0.039/0.041	0.239/0.234	0.034/0.036	0.239/0.234
EDQA (Ours)	**0.132/0.129**	**0.423/0.422**	**0.180/0.170**	**0.420/0.415**	**0.207/0.207**	**0.466/0.463**	**0.176/0.172**	**0.420/0.413**

TABLE VII

RESULTS OF QUALITY SCORE REGRESSION WITH SRCC / PLCC METRICS. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

(a) Full-reference setting.

Methods	PIPAL^{test}	KADID^{test}	TID2013	CSIQ
SSIM	0.624 / 0.680	0.750 / 0.751	0.746 / 0.802	0.861 / 0.857
FSIM	0.673 / 0.746	0.855 / 0.857	0.841 / 0.875	**0.937** / 0.937
LPIPS	0.639 / 0.718	0.799 / 0.803	0.798 / 0.851	0.905 / 0.926
EDQA (Ours)	**0.743 / 0.780**	**0.938 / 0.943**	**0.852 / 0.886**	0.934 / **0.949**

(b) Non-reference setting.

Methods	PIPAL^{test}	KADID^{test}	TID2013	CSIQ
NIQE	0.300 / 0.367	0.430 / 0.499	0.315 / 0.413	0.660 / 0.747
CLIPIQA	0.448 / 0.491	0.644 / 0.653	0.616 / 0.690	0.761 / 0.798
MUSIQ	0.539 / 0.570	0.650 / 0.668	0.578 / 0.693	0.755 / 0.811
MANIQA	0.558 / 0.602	0.482 / 0.527	0.472 / 0.603	0.701 / 0.714
EDQA (Ours)	**0.742 / 0.778**	**0.937 / 0.941**	**0.847 / 0.866**	**0.912 / 0.938**

TABLE VIII

RESULTS OF QUALITY SCORE CALCULATION ON SPAQ AND KONIQ DATASETS WITH SRCC / PLCC METRICS. EDQA NEEDS TO COMPARE IMAGES WITH DIFFERENT CONTENTS TO OBTAIN THE QUALITY SCORE, SINCE ALL IMAGES IN THE TWO DATASETS CONTAIN DIFFERENT CONTENTS.THE ORIGINAL EDQA IS ONLY TRAINED TO COMPARE IMAGES WITH SIMILAR CONTENTS, WHICH BRINGS A TASK GAP.WHEN TRAINED ON THE SAME DATASET AS BASELINES, EDQA SURPASSES THE BASELINE METHODS

(a) Results on SPAQ dataset.

Methods	NIQE	CLIPIQA	MUSIQ	MANIQA	EDQA	EDQA
Train Set	-	-	KonIQ	KonIQ	KonIQ	Original
SRCC	0.664	0.700	0.856	0.755	**0.859**	0.835
PLCC	0.679	0.722	0.859	0.765	**0.861**	0.841

(b) Results on KonIQ dataset.

Methods	NIQE	CLIPIQA	DBCNN	MUSIQ	EDQA	EDQA
Train Set	-	-	SPAQ	SPAQ	SPAQ	Original
SRCC	0.530	0.685	0.731	0.753	**0.787**	0.717
PLCC	0.533	0.717	0.758	0.680	**0.807**	0.729

TABLE IX

OUR ASSISTANT MODEL SURPASSES GPT-4V IN INSTANT RATING TASK. THE METRIC IS ACCURACY IN THE FULL-REFERENCE / NON-REFERENCE SETTING. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	TID2013	LIVE-MD	MDID2013
GPT-4V	84.4 / 88.1	79.6 / 72.7	70.6 / 67.6
Our Assistant Model	**94.9 / 94.6**	**93.1 / 92.8**	**90.1 / 89.8**

implement a trainable image abstractor, which is a four-layer transformer network, to map visual tokens into the textual space following [3]. The abstractor can also significantly

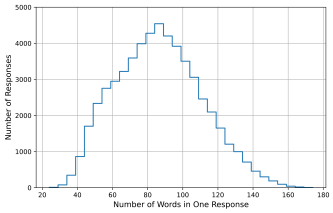


Fig. 5. **Word length distribution** of detailed responses.

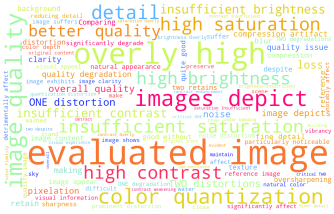


Fig. 6. **Wordcloud** map of all detail responses.

reduce the number of vision tokens, relieving the computing pressure. (2) Token fusion. We integrate the visual tokens into pre-defined positions within the textual tokens as token fusion. (3) Response generation using LLM. The fused tokens are fed into LLM to generate the final response. Here we mainly conduct experiments with Vicuna-v1.5-7B. Despite their capabilities, pre-trained LLMs typically do not perform well on IQA tasks without adjustments. Therefore, we employ LoRA [90], a fine-tuning technique that efficiently modifies a small subset of parameters within the LLM. Specifically, we apply LoRA to adjust the projection layers in all self-attention modules, following [49], [90]. This approach allows for targeted refinement of the model's performance on IQA tasks without the need for extensive retraining. Our experiments in Tab. XIII also show that model architecture has little influence on model performance.

B. Retaining Resolution in Training

Although previous VLM-based IQA models typically resize all input images to a fixed resolution [5], [6], we find this might hurt their performance, as resolution variation may affect visual quality. Instead, we retain the original image resolution during training. Specifically, we interpolate (in bicubic mode) the position embedding in CLIP [83] to accommodate varying image resolutions. Ablation studies detailed in Sec. V-M demonstrate our model's capability to assess quality variations attributable to resolution, even without explicitly training on such tasks.

TABLE X

RELATIONSHIPS BETWEEN THE COMPARISON REASONING AND INSTANT RATING TASKS. OVERALL, THE TWO TASKS ARE BENEFICIAL TO EACH OTHER

(a) Results on the instant rating task.

	BAPPS^{test}	KADID^{test}	PIPAL^{test}	TID2013	LIVE-MD	MDID2013
Only Rating	81.6 / 81.6	92.4 / 92.3	89.1 / 89.0	94.2 / 94.1	**92.9** / **92.7**	**92.1** / **91.7**
Co-training	**84.7** / **82.4**	**93.6** / **93.1**	**90.5** / **90.0**	**96.9** / **96.4**	92.1 / 91.8	90.0 / 89.6

(b) Results on the comparison reasoning task.

	Single-distortion			Multi-distortion		
	GPT-4 Score	BLEU	ROUGE-L	GPT-4 Score	BLEU	ROUGE-L
Only Reasoning	74.3 / 69.6	0.203 / 0.202	0.465 / 0.453	70.6 / **69.1**	0.165 / 0.165	0.414 / 0.407
Co-training	**75.1** / **74.9**	**0.207** / **0.207**	**0.466** / **0.463**	**71.7** / 68.7	**0.176** / **0.172**	**0.420** / **0.413**

TABLE XI

RESULTS ON Q-BENCH [4]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Q-Bench	Yes-or-No	What	How	Distortion	Other
LLaVA (Q-Instruct trained)	66.4	58.2	50.5	49.4	65.7
EDQA (Q-Instruct trained)	70.5	49.6	51.7	55.4	56.7
EDQA (Q-Instruct + EDQA trained)	**74.0**	**62.8**	**54.0**	**66.7**	**63.9**

TABLE XII

RETAINING RESOLUTION DURING BOTH TRAINING AND INFERENCE IS IMPORTANT TO IDENTIFY IMAGES WITH BETTER ASPECT RATIO OR HIGHER RESOLUTION

Retain Resolution? Training	Inference	H↔W	0.5×	0.75×	0.8×	0.85×	0.9×	0.95×
✗	✗	73.0	99.0	93.5	91.7	83.8	77.2	71.2
✓	✗	85.6	99.8	99.4	99.0	95.9	94.8	89.4
✓	✓	98.8	99.9	99.6	99.3	99.1	96.8	97.0

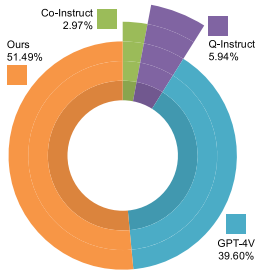


Fig. 7. User study results.

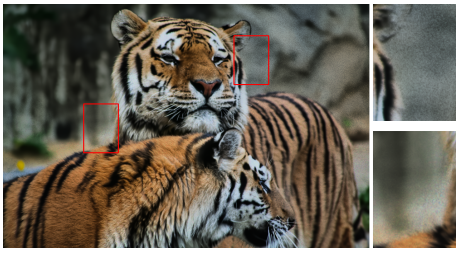


Fig. 8. One example of model-restored image by SwinIR [100].

C. Confidence Estimation

In many applications, it is important to know a confidence score that indicates when the model is uncertain of its response. Here we use the confidence scores of some key tokens as the confidence of the entire answer. Intuitively, the key tokens are distortion names in *distortion identification*, and are either "Image A" or "Image B" in *instant rating*. For detailed reasoning tasks, which feature diverse and non-structured responses, we utilize semantic change testing [91] to identify the top 20 tokens with the highest importance scores as key tokens. In semantic change testing, we employ all-MiniLM-L6-v2 [92] as the similarity model, due to its high processing speed (14K sentences per second). The predicted likelihood of key tokens is averaged as the confidence score. Results in Fig. 9 verify that confidence and model performance are highly correlated.

D. Model Setup

Since the CLIP pre-trained ViT-L/14 [83] encodes each 14×14 patch to a visual token, the resolution of the input image should be integer multiples of 14. Therefore, we first pad the size of input images to integer multiples of 14 with zero-padding. We encode the image patches into visual tokens using the CLIP pre-trained ViT-L/14 [83], with each token having a channel of 1024. The vision abstractor can reduce the number of vision tokens to 64 and map the vision tokens to the hidden dimension of the LLM, which is 4096. Without the vision abstractor, the maximum resolution is limited to 672, constrained by computation resources (RTX A6000 GPUs). However, with the vision abstractor, we can process images with much larger resolutions (up to 2500×2500). In our experiments, the maximum image resolution is 1092×1456, thus the resolutions of all images are retained. The vision abstractor consists of four transformer layers with 64 learnable query embeddings. In LoRA of LLM, the parameters of rank and scale factor are both set as 16. The projection weights in each attention layer are fine-tuned using LoRA technique.

E. Training and Inference Setup

EDQA is trained for 1 epoch with batch size 64. Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.001, and learning rate 0.0002 is used for training. During inference, the temperature is set to 0, since lots of predicted information (*e.g.*, distortion, comparison result) need to be certain.

V. EXPERIMENTS

A. Metrics and Baselines

B. Accuracy, SRCC, and PLCC

The accuracy metric is utilized for *distortion identification* and *instant rating* tasks. VLMs usually produce diverse textual

TABLE XIII
SIMPLY REPLACING MODULES} (I.E., VISION-TEXT CONNECTORS AND LLMS) HAS RELATIVELY LITTLE INFLUENCE ON PERFORMANCE

Types	Architectures	Distortion Identification		Instant Rating					
		Single-dist. ID	Multi-dist. ID	BAPPS^{test}	KADID^{test}	PIPAL^{test}	TID2013	LIVE-MD	MDID2013
Vision-text connector	Projector	97.9 / 94.7	90.5 / 89.5	82.7 / 81.4	92.7 / 92.4	89.2 / 88.8	96.2 / 95.9	92.1 / 91.9	89.1 / 88.4
	Abstractor	97.7 / 94.1	91.3 / 89.3	84.7 / 82.4	93.6 / 93.1	90.5 / 90.0	96.9 / 96.4	92.1 / 91.8	90.0 / 89.6
LLM	Vicuna-v1.5-7B	97.9 / 94.7	90.5 / 89.5	82.7 / 81.4	92.7 / 92.4	89.2 / 88.8	96.2 / 95.9	92.1 / 91.9	89.1 / 88.4
	Vicuna-v0-7B	96.9 / 93.6	89.8 / 89.3	82.5 / 81.3	92.8 / 92.2	88.2 / 88.1	95.0 / 94.7	91.2 / 91.1	90.5 / 90.2
	LLaMA-2-7B	97.0 / 94.0	90.6 / 89.1	81.7 / 81.5	92.6 / 92.0	88.4 / 87.9	94.6 / 94.1	91.8 / 91.1	90.9 / 90.7

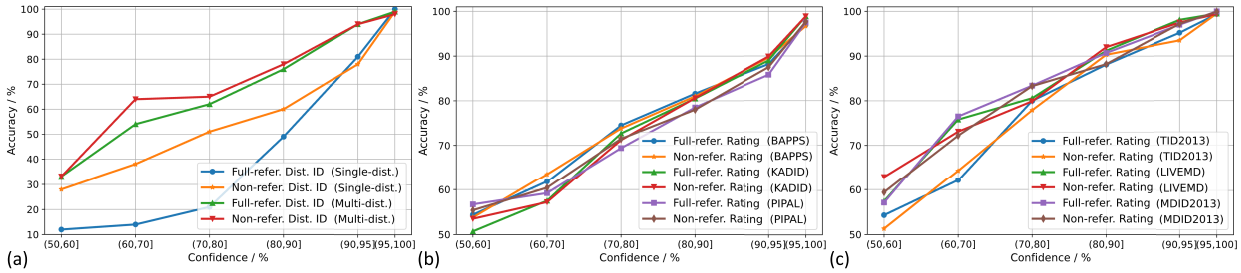


Fig. 9. **Our estimated confidence scores** are high correlated to the model performance on(a) *distortion identification* and (b) (c) *instant rating* tasks on different benchmarks in both full-reference and non-reference settings.

outputs, and we transform them into brief results for accuracy calculation. Specifically, we prompt our EDQA with "Answer the question using a single word or phrase" to encourage direct output of brief responses. For baseline models, we include all potential answers in the prompt and instruct the model to identify the most accurate one. We emphasize that *our key motivation is to generate descriptive language rather than quality scores*. However, our approach can produce quality scores using pair-wise comparison if required. The quality scores are assessed using Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC).

C. GPT-4 score, BLEU, and ROUGE-L

We employ the GPT-4 score to evaluate *assessment reasoning* and *comparison reasoning* tasks, following [1]. Specifically, we provide GPT-4 with both the model-generated response and the corresponding ground truth response. GPT-4 assesses the helpfulness, relevance, accuracy, and level of detail in the model-generated response relative to the ground truth, assigning an overall score on a scale of 0 to 10, where a higher score indicates better quality. This average score is subsequently normalized to a scale of 0 to 100%, reported as the GPT-4 score metric. We further evaluate the reasoning tasks with classical metrics including BLEU and ROUGE-L score following [93], [94].

D. Baselines

We categorize our baseline methods into general-purpose VLMs and IQA-specific VLMs.

Note that Q-Instruct only supports single-image inputs, thus we only test it on non-reference single-image assessment tasks. Additionally, we compare traditional score-based IQA methods including full-reference ones (PSNR, SSIM [14], LPIPS [10], DISTS [19]) and non-reference ones (NIQE [26], ClipIQA [38], MUSIQ [11], MANIQA [39]) in *instant rating* task and score regression experiments.

E. Results on Benchmarks

Quantitative results of distortion identification are shown in Tab. III.

Fourth, EDQA stably surpasses all baseline methods, demonstrating our model's efficacy. Finally, we evaluate our model in an out-of-distribution (OOD) setting, *i.e.*, for a particular category of distortion (*e.g.*, noise), we use some sub-categories (*e.g.*, Gaussian noise) during training, and other sub-categories (*e.g.*, impulse noise) for evaluation. Results in the last column of Tab. III show that our method maintains high accuracy even under such an OOD setting.

Quantitative results of instant rating are demonstrated in Tab. IV. First, in the full-reference context, traditional score-based methods, even the simplest PSNR, outperform all general VLMs including GPT-4V and prior IQA-specific VLMs, indicating the inadequacy of existing VLMs in full-reference IQA tasks. Second, conversely, in the non-reference scenario, GPT-4V and Co-Instruct excel beyond most score-based approaches, except MUSIQ.

Finally, EDQA demonstrates superior performance across both settings by a large margin, showcasing its substantial advantage.

Quantitative results of assessment reasoning and comparison reasoning are shown in Tab. V and Tab. VI. First, the performance of the VLM-specific models significantly declines on tasks outside their defined scopes. For instance, Co-Instruct's performance is unsatisfactory on full-reference tasks. Second, GPT-4V shows robust reasoning abilities, stably outperforming prior IQA-specific VLMs. Third, EDQA surpasses GPT-4V, especially in the non-reference setting, affirming its superior reasoning abilities. Finally, EDQA achieves relatively good GPT-4 score and ROUGE-L,

indicative of the overall semantic accuracy, but a low BLEU score (yet remains much higher than GPT-4V), which reflects word-level consistency. This suggests that while our predicted answers do not precisely duplicate the ground truths word-for-word, they preserve similar meanings with diverse expressions.

Qualitative results of our model on the four tasks in the non-reference setting are depicted in Fig. 2. More qualitative results are provided in *Supp. Mat.*

F. Quality Score Regression

Our key focus in this work is to *generate descriptive language rather than quality scores*. We focus more on linguistic descriptions because language is an effective interaction tool in an LLM-based intelligent agent. With the rapid development of LLMs and multi-modal techniques, in an LLM-based intelligent agent, language could be a useful tool for interacting and communicating across quality-related tasks such as image assessment, refinement, editing, and recommendation. Still, if it is required, our approach can produce quality scores.

G. Quality Score Regression

The score regression results are evaluated on the PIPAL, KADID, TID2013, and CSIQ datasets. These datasets include high-quality reference images and their distorted versions under various distortions. We calculate the win rate of an image against others to determine its quality score. Specifically, for an image A, we randomly sample comparison candidates, such as B, C, D, *etc.*, which share the same content as A but have different distortions. Image A is then compared pairwise with each of its comparison candidates (B, C, D, *etc.*). Finally, the win rate of Image A against all its compared candidates is calculated as its quality score. The comparison numbers per image for PIPAL, KADID, TID2013, and CSIQ datasets are 58, 62, 60, and 15, respectively. We show that the comparison number per image could be reduced significantly without large performance degradation in *Supp. Mat.* The results of quality score regression are given in Tab. VII and Tab. VII, proving that our method can generate accurate quality scores.

H. Assessing in-The-Wild Images With Different Contents

Existing real-world IQA datasets like KonIQ [98] and SPAQ [99] contain real distorted images with various contents. To regress quality scores from such a dataset, our model needs to compare images with different contents though *it is trained only to compare images with similar contents*, as shown in Task 2 of Fig. 2. The results in Tab. VIII show that even with a task gap between training and test, our original EDQA still achieves comparable results with previous score-based IQA methods in score regression. Furthermore, we formulate real-world IQA datasets into instant rating tasks to re-train our EDQA, *i.e.*, trained on KonIQ then evaluated on SPAQ, and vice versa. Our re-trained EDQA outperforms all baselines trained on the same dataset. These results indicate that our method is capable of assessing in-the-wild images with different contents.

I. Ablation Studies

Relationships between the comparison reasoning and instant rating tasks are studied in Tab. X. First, comparison reasoning task improves the performance on four instant rating datasets, but decreases the results on two datasets. Overall, comparison reasoning task helps the instant rating. Second, instant rating task stably improves the performance on comparison reasoning task.

J. Confidence

We examine the correlation between the performance and estimated confidence in Fig. 9. For *distortion identification* and *instant rating* tasks, across both full-reference and non-reference settings, our model demonstrates improved performance as the confidence interval increases. This validates the effectiveness of our confidence estimation.

K. Assistant Model

To construct *comparison reasoning* responses, we train an assistant model to predict comparison results (see Fig. 3). These results serve as pseudo labels, which are subsequently provided to GPT-4V to generate responses. We compare the assistant model to GPT-4V on three out-of-distribution IQA datasets. The results in Tab. IX affirm the superiority of the assistant model.

L. Retaining Resolution

In Tab. XII, we study the effects of retaining resolution. Specially, we randomly sample 1,000 high-quality images whose aspect ratios are greater than 4:3. These images are either resized by swapping their height and width (denoted as H↔W), or bi-linearly down-sampled by a scale factor of 0.5, 0.75, 0.8, 0.85, 0.9, or 0.95.EDQA is requested to conduct the *instant rating* task, *i.e.*, compare the original and resized images to determine the superior one. The alternative method of retaining resolution is to resize both original image and resized image to a larger resolution, which can maintain the quality difference. In contrast, resizing both images to smaller resolution results in two nearly same images. The results are presented in Tab. XII. First, retaining resolution is crucial for identifying images with better aspect ratio or higher resolution. Second, with down-sampling becomes severer (*i.e.*, aspect ratio is from 0.95 to 0.5), the accuracy is improved since the quality drop is more significant. Third, for severe down-sampling (*e.g.*, aspect ratio is 0.5) where the quality degradation is quite obviously, retaining resolution or just resizing both images to a larger size both perform well (

). Finally, however, for relatively slight down-sampling (*e.g.*, aspect ratio is from 0.75 to 0.95), the performance of retaining resolution is stably superior than resizing.

Considering that vision abstractor can greatly reduce the computational burden than projector, we select abstractor by default. For example, for a 448×448 image, projector generates 1024 tokens, while abstractor only outputs 64 tokens. Note that the computation amount is proportional to the square of the number of tokens.

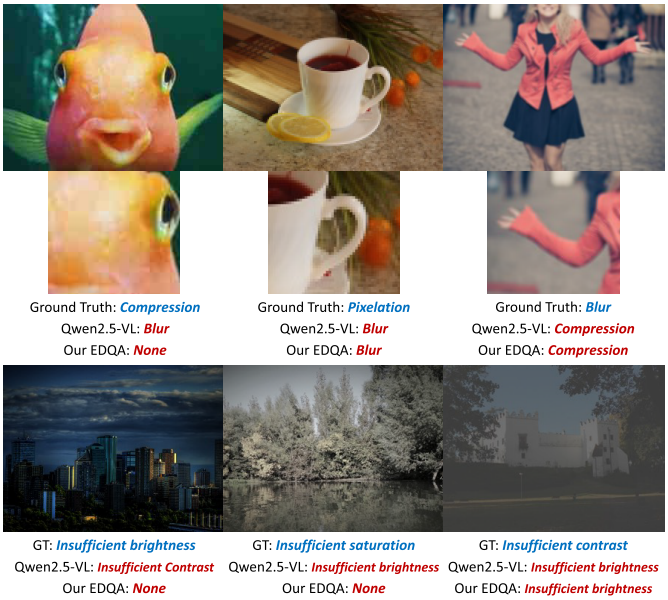


Fig. 10. Failure cases where the model confuses some similar distortions, leading to misclassification.

TABLE XIV

ABLATION RESULTS OF COMPARISON ORDER ON FINE-GRAINED DATASET RELEASED IN [101]. WE FOLLOW [101] TO REPORT THE CONSISTENCY (%) / ACCURACY (%) / CORRELATION AS METRICS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

	Q-Instruct [66]	GPT-4V [2]	EDQA (Ours)
CSIQ (Levels)	11.5 / 8.1 / 0.557	41.9 / 40.2 / 0.906	**95.5 / 92.5 / 0.958**
CSIQ (Types)	11.7 / 6.9 / 0.416	32.5 / 24.4 / 0.482	**90.5 / 69.0 / 0.857**
SPAQ (Scores)	44.8 / 23.3 / 0.328	65.3 / 39.8 / 0.448	**92.1 / 59.6 / 0.961**

TABLE XV

CONFIDENCE ESTIMATION OF CONSISTENT AND INCONSISTENT COMPARISON RESULTS ON FINE-GRAINED DATASET IN [101]

Confidence statistics	Consistency	Inconsistency
CSIQ (Levels)	0.933 ± 0.120	0.629 ± 0.083
CSIQ (Types)	0.860 ± 0.141	0.611 ± 0.098
SPAQ (Scores)	0.900 ± 0.125	0.649 ± 0.110

TABLE XVI

RESULTS ON MODEL-PROCESSED IMAGES. THE BEST RESULTS ARE BOLD. GPT-4V EXHIBITS VARIABILITY IN RANKING, SO ITS RESULTS ARE REPORTED AS MEAN ± STANDARD DEVIATION

	NIQE	ClipIQA	MUSIQ	ManIQA	GPT-4V	EDQA (Ours)
Rank ↓	2.20	1.40	1.60	1.80	1.34±0.27	**1.20**
Accuracy ↑	45.5	72.7	77.3	66.4	74.5	**82.7**

TABLE XVII

IMAGE RESTORATION MODELS USED, INCLUDING SWINIR [100], HAT [102], X-RESTORMER [103], MPRNET [104], RESTORMER [105], FBCNN [106], MAXIM [107], MPRNET [104], DRBNET [108], AND IFAN [109]. FOR FBCNN, "q = 90" MEANS TRAINING ON THE QUALITY FACTOR 90, AND "BLIND" MEANS BLIND TO THE QUALITY FACTOR

Image restoration task	Image restoration models
Super-resolution	SwinIR, HAT, X-Restormer
Denoising	SwinIR, MPRNet, Restormer, X-Restormer
JPEG artifact removal	SwinIR, FBCNN (q=90), FBCNN (blind)
Motion deblurring	MAXIM, MPRNet, Restormer
Defocus deblurring	DRBNet, IFAN, Restormer

M. Real-World Applications

N. Assessing Web-Downloaded Images

A practical usage of an IQA model involves assessing the quality of real images. We collect a total of 50 real-world images from the web, featuring diverse contents including animals, plants, faces, buildings, and landscapes. Qualitative results in Fig. 11 indicate that our method can assess real images with detailed descriptions. More importantly, EDQA can describe how the distortions affect the contents. For example, in Fig. 11d, our model first accurately identifies the "severe quantization", then describes that the quantization "causes banding in the sky and water", and finally concludes that the quality "is considerably degraded". We also conduct a user study with 20 participants involved. Participants are instructed to choose the assessment result that is of the highest quality among the test methods. The results are shown in Fig. 7, revealing that our approach stably outperforms baseline methods in aligning human perception.

O. Comparison on Model-Processed Images

To develop image restoration models, one often needs to compare the restoration quality of different models. We consider five common image restoration tasks: super-resolution, denoising, JPEG compression artifact removal, motion deblurring, and defocus deblurring. For each task, we collect three to four cutting-edged models in recent years (listed in Tab. XVII), apply them to a correspondingly degraded image, and then manually rank the resultant model-processed images. To find the image considered best by VLMs, we linearly scan the candidates and compare them in pairs. As VLMs' results are not deterministic and may be sensitive to the presentation order of images, we repeat the linear scan 10 times and randomly shuffle the scan order each time. The average rank of the found best restoration and the accuracy of the paired comparison are reported in Tab. XVI. First, EDQA achieves an average rank of 1.20 (1 is the best), outperforming both GPT-4V and score-based methods. Second, though the temperature is set to 0, GPT-4V shows variability with a large standard deviation. Third, model-restored images are generally out-of-distribution for our model, while EDQA exhibits excellent generalization ability on these images. For example, the image in Fig. 8 is restored from a noisy image. There is still remnant noise, which is somewhat strange. For such an OOD image, our EDQA correctly recognizes it to be inferior, but MANIQA, MUSIQ, and NIQE consider it as the best of the four candidates. Finally, we provide two qualitative results of detailed comparison reasoning on model-processed images in Fig. 12. One compares SwinIR and FBCNN in the JPEG compression artifact removal task, and the other compares MPRNet and MAXIM in the deblur task. Our model can generate a reasonable explanation for the comparison results.

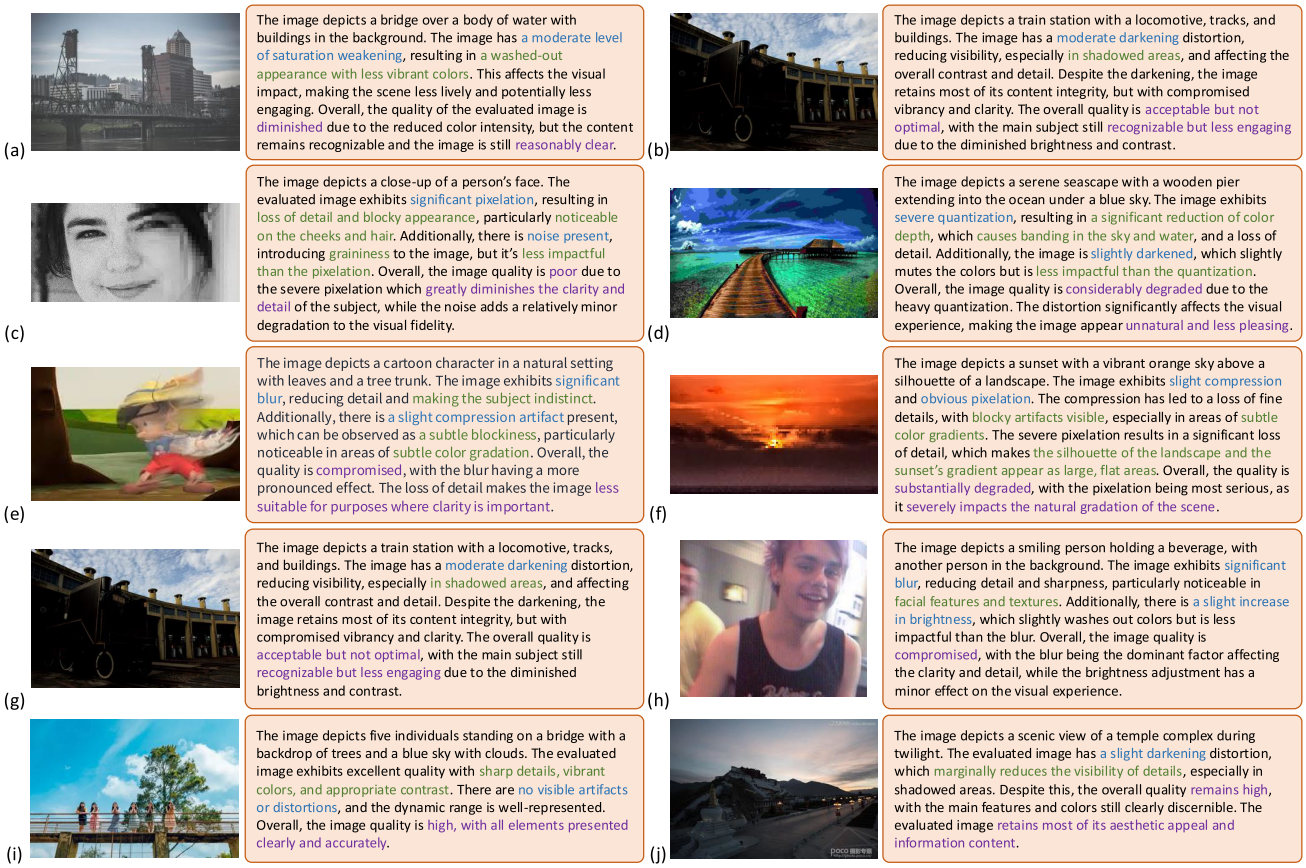


Fig. 11. **Qualitative results** on assessing web-downloaded images. More results in Fig. S6 of *Supp. Mat*.

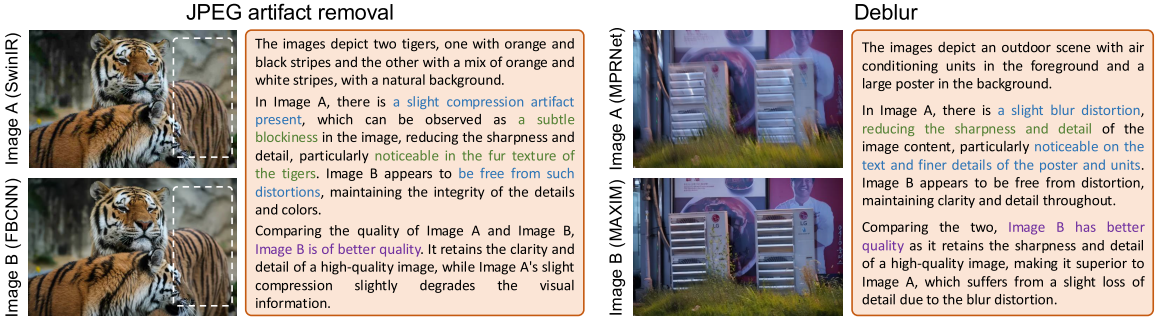


Fig. 12. **Qualitative results** of detailed comparison reasoning on model-processed images.

P. Complexity and Efficiency

Q. Training Cost

The total parameters are 7.11B, including 6.76B for LLM, 0.30B for vision encoder, and 54M for vision abstractor. The trainable parameters are 70M (54M for vision abstractor and 16M for LoRA), constituting only 0.98% of the total parameters. The model is trained on 8 GPUs (RTX A6000). The training is completed in around 22 hours.

R. Inference Cost

The inference latency depends on the response length and it is tested on a single RTX A6000 GPU. For brief tasks task with the short answer prompt (about 2.92 words), the inference time stands at approximately 2.23s / batch=32, transformed to 0.07s / sample. For the assessment reasoning task (75.84 words on average), the inference time is 22.97s / batch=32 (*i.e.*, 0.72s per response). EDQA remains deployable on a single consumer GPU (*e.g.*, RTX3090).

VI. Conclusion and Limitations

We introduce EDQA, a VLM-based IQA model, empowered by a new multi-functional task paradigm, dataset enrichment, and training technique, surpassing baseline methods in both benchmarks and two real-world applications, showing the potential of descriptive quality assessment.

*Limitations:*First, the fine-grained abilities requiring more high-level perception skills are still unsatisfactory.

For example, in Fig. 11c, though identifying noise and pixelation successfully, our model fails to point out that they are respectively located in the left and right parts. One solution is to take the segmentation model to add various distortions to different regions. Second, for the convenience of evaluating, analyzing, and improving the model, we mainly focus on standardized answers. To achieve more flexible responses, LLM rewriting and human annotation can be used to increase linguistic diversity during dataset construction. Third, whether our assessment can be used as feedback to improve the quality of generation or restoration models is still under-explored.

REFERENCES

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2023, pp. 34892–34916.

[2] OpenAI.(2023). *GPT-4V(Ision) System Card*. [Online]. Available: https://openai.com/research/gpt-4v-system-card

[3] Q. Ye et al., "MPLUG-OwI2: Revolutionizing multi-modal large language model with modality collaboration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13040–13051.

[4] H. Wu et al., "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *Proc. Int. Conf. Learn. Represent.*, 2023.

[5] H. Wu et al., "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25490–25500.

[6] H. Wu et al., "Towards open-ended visual quality comparison," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 360–377.

[7] T. Wu, K. Ma, J. Liang, Y. Yang, and L. Zhang, "A comprehensive study of multimodal large language models for image quality assessment," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 143–160.

[8] Z. You, Y. Li, J. Gu, Z. Yin, T. Xue, and C. Dong, "Depicting beyond scores: Advancing image quality assessment through multi-modal language models," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 259–276.

[9] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[11] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5148–5157.

[12] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 633–651.

[13] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[15] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[16] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[17] 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 Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[18] Y. Cao, Z. Wan, D. Ren, Z. Yan, and W. Zuo, "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5851–5861.

[19] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[20] K. Ding, Y. Liu, X. Zou, S. Wang, and K. Ma, "Locally adaptive structure and texture similarity for image quality assessment," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2483–2491.

[21] A. Ghildyal and F. Liu, "Shift-tolerant perceptual similarity metric," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 91–107.

[22] G. Yin et al., "Content-variant reference image quality assessment via knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 3134–3142.

[23] W. Zhou and Z. Wang, "Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 934–942.

[24] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.

[25] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[26] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2012.

[27] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[28] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[29] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[30] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 305–312.

[31] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.

[32] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

[33] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6373–6382.

[34] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3664–3673.

[35] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 25, pp. 2912–2925, 2023.

[36] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10222–10231.

[37] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14143–14152.

[38] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, 2023, pp. 2555–2563.

[39] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1191–1200.

[40] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2864–2878, Mar. 2023.

[41] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14071–14081.

[42] W.-L. Chiang et al. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90% ChatGPT Quality*. [Online]. Available: https://vicuna.lmsys.org

[43] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[44] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[45] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.

[46] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2023, pp. 49250–49267.

[47] H. Wei et al., "Vary: Scaling up the vision vocabulary for large vision-language model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 408–424.

[48] Q. Ye et al., "MPLUG-owl: Modularization empowers large language models with multimodality," 2023, *arXiv:2304.14178*.

[49] Z. Yin et al., "LAMM: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2023, pp. 26650–26685.

[50] P. Zhang et al., "InternLM-XComposer: A vision-language large model for advanced text-image comprehension and composition," 2023, *arXiv:2309.15112*.

[51] R. Zhang et al., "LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention," in *Proc. Int. Conf. Learn. Represent.*, 2024, pp. 1–8.

[52] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *Proc. Int. Conf. Learn. Represent.*, 2023.

[53] H. Agrawal et al., "Nocaps: Novel object captioning at scale," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8948–8957.

[54] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[55] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[56] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[57] Y. Liu et al., "MMBench: Is your multi-modal model an all-around player?," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 216–233.

[58] P. Lu et al., "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 2507–2521.

[59] A. Masry, D. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 2263–2279.

[60] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2200–2209.

[61] A. Singh et al., "Towards VQA models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8309–8318.

[62] H. Zhu et al., "2AFC prompting of large multimodal models for image quality assessment," 2024, *arXiv:2402.01162*.

[63] H. Wu et al., "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 54015–54029.

[64] Z. You, X. Cai, J. Gu, T. Xue, and C. Dong, "Teaching large language models to regress accurate image quality scores using score distribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 14483–14494.

[65] D. Guo et al., "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.

[66] W. Li et al., "Q-insight: Understanding image quality via visual reinforcement learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2025, pp. 1–11.

[67] Z. Shao et al., "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," 2024, *arXiv:2402.03300*.

[68] T. Wu, J. Zou, J. Liang, L. Zhang, and K. Ma, "VisualQuality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2025.

[69] Z. Cai et al., "Q-ponder: A unified training pipeline for reasoning-based visual quality assessment," 2025, *arXiv:2506.05384*.

[70] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[71] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

[72] Y. Hao, L. Dong, F. Wei, and K. Xu, "Self-attention attribution: Interpreting information interactions inside transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 14, pp. 12963–12971.

[73] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs," 2020, *arXiv:2008.02312*.

[74] P. Schramowski et al., "Making deep neural networks right for the right scientific reasons by interacting with their explanations," *Nature Mach. Intell.*, vol. 2, no. 8, pp. 476–486, Aug. 2020.

[75] Y. Shen et al., "Domain-invariant interpretable fundus image quality assessment," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101654.

[76] B. Jo, D. Cho, I. Kyu Park, and S. Hong, "IFQA: Interpretable face quality assessment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3444–3453.

[77] P. Costa et al., "EyeQual: Accurate, explainable, retinal image quality assessment," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 323–330.

[78] Y. Hu et al., "TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20349–20360.

[79] H. Lin, V. Hosu, and D. Saupe, "DeepFL-IQA: Weak supervision for deep IQA feature learning," 2020, *arXiv:2001.08113*.

[80] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[81] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Conf. Rec. Forty 6th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 1693–1697.

[82] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[83] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[84] R. A. Frazor and W. S. Geisler, "Local luminance and contrast in natural images," *Vis. Res.*, vol. 46, no. 10, pp. 1585–1598, May 2006.

[85] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[86] P. J. Bex, S. G. Solomon, and S. C. Dakin, "Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure," *J. Vis.*, vol. 9, no. 10, pp. 1–19, Sep. 2009.

[87] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, vol. 39. Cham, Switzerland: Springer, 2009.

[88] H. Liu et al. (2024). *LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge*. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[89] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2018, pp. 66–71.

[90] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[91] J. Duan et al., "Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 5050–5063.

[92] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[93] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1073–1083.

[94] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.

[95] X. An et al., "LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training," 2025, *arXiv:2509.23661*.

[96] Z. Chen et al., "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2024, *arXiv:2412.05271*.

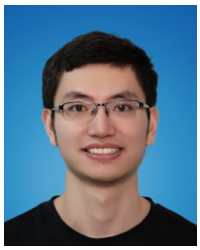
[97] S. Bai et al., "Qwen2.5-VL technical report," 2025, *arXiv:2502.13923*.

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

[99] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2020, pp. 3677–3686.
[100] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[101] H. Zhu et al., "2AFC prompting of large multimodal models for image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 12, pp. 12873–12878, Dec. 2024.
[102] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.
[103] X. Chen et al., "A comparative study of image restoration networks for general backbone network design," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 74–91.
[104] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.
[105] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.
[106] J. Jiang, K. Zhang, and R. Timofte, "Towards flexible blind JPEG artifacts removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4977–4986.
[107] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5769–5780.
[108] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16283–16292.
[109] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2034–2042.

Zhiyuan You received the bachelor's and master's degrees from Shanghai Jiao Tong University, where he was supervised by Prof. Xinyi Le and Prof. Yu Zheng. He is currently pursuing the Ph.D. degree with the Multimedia Laboratory, The Chinese University of Hong Kong, under the supervision of Prof. Chao Dong and Prof. Tianfan Xue. His research interests focus on low-level vision based on foundation models.

Jinjin Gu received the Ph.D. degree from the School of Electrical and Computer Engineering, The University of Sydney in 2024. He is currently a tenure-track Faculty Member with INSAIT, Sofia, Bulgaria. His research contributions have received over 13,000 citations, achieving an H-index of 35. His research interests include visual perception, visual processing, visual generation, and visual reasoning. In 2024, he was named among the top 2% of scientists worldwide by Stanford University. He received the Yunfan Award at the World Artificial Intelligence Conference (WAIC) in 2023.

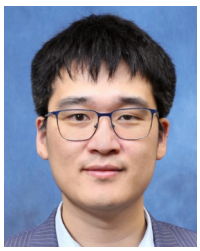


Xin Cai received the B.S. degree in computer science from the University of Chinese Academy of Sciences (UCAS) in 2020 and the M.S. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. He is currently pursuing the Ph.D. degree with The Chinese University of Hong Kong (CUHK). His research interests include computer vision, computational photography, and machine learning.

Zheyuan Li received the B.E. degree from Northwestern Polytechnical University, Xi'an, in 2022. He is currently pursuing the Ph.D. degree with the University of Macau. He has been a Research Intern with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences since 2021. His research interests include image super-resolution, image restoration and enhancement, and multi-modal low-level-vision model.

Kaiwen Zhu received the B.Eng. degree from Shanghai Jiao Tong University in 2024, where he is currently pursuing the Ph.D. degree, under the supervised by Prof. Chao Dong. He has been an Intern Researcher with Shanghai Artificial Intelligence Laboratory since 2023. His research focuses on low-level vision and intelligent agents.

Chao Dong is currently a Professor with Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science, and Shanghai Artificial Intelligence Laboratory. In 2014, he first introduced deep learning method–SRCNN into the super-resolution field. This seminal work was chosen as one of the top ten "Most Popular Articles" of TPAMI in 2016. His current research interest focuses on low-level vision problems, such as image/video super-resolution, denoising, and enhancement. His team has won several championships in international challenges–NTIRE2018, PIRM2018, NTIRE2019, NTIRE2020, AIM2020, and NTIRE2022. He worked in Sense Time from 2016 to 2018, as the Team Leader of Super-Resolution Group. In 2021, he was chosen as one of the World's Top 2% Scientists. In 2022, he was recognized as the Al 2000 Most Influential Scholar Honorable Mention in computer vision.



Tianfan Xue received the bachelor's degree from Tsinghua University, the M.Phil. degree from CUHK in 2011, and the Ph.D. degree from the Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT in 2017. He is currently a Vice-Chancellor Assistant Professor with the Multimedia Laboratory, Department of Information Engineering, The Chinese University of Hong Kong (CUHK). Prior to this, he was a Google Researcher with the Computational Photography Team, for over five years. His research focuses on computational photography, 3D reconstruction, and generation. His work on bilateral based 3D reconstruction has won SIGGRAPH Honorable mention 2024. He also served as an Area Chair for WACV, CVPR, NeurIPS, and ACM MM.