# ENHANCING DESCRIPTIVE IMAGE QUALITY ASSESS MENT WITH A LARGE-SCALE MULTI-MODAL DATASET

Anonymous authors

004

006 007 008

009 010

011

012

013

014

015

016

017

018

019

021

023

025

026

027

028

029

031

Paper under double-blind review

### ABSTRACT

With the rapid advancement of Vision Language Models (VLMs), VLM-based Image Quality Assessment (IOA) 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 realworld applications. Second, their performance is sub-optimal due to limitations in dataset coverage, scale, and quality. To overcome these challenges, we introduce Enhanced Descriptive image Quality Assessment (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 realworld applications including assessing the web-downloaded images and ranking model-processed images. Datasets and codes will be released publicly.

### 1 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) (Liu et al., 2023a; OpenAI, 2023; Ye et al., 2023b), VLM-based IQA begins to attract more research interest (Wu et al., 2024a;b;d;e; You et al., 2023). 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 (You et al., 2023). 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 040 only support a few of them. For example, one scenario involves assessing a single image downloaded 041 from the web, while another requires comparing multiple images handled by different algorithms. 042 Also, image restoration needs to assess an image against a reference, while image generation requests 043 non-reference assessments. Therefore, a superior IQA model should be multi-functional to cater 044 to such diverse scenarios. However, existing methods limit to a specific subset of these tasks, such as single-image assessment (Wu et al., 2024b), multi-image comparison (Wu et al., 2024d), or full-reference setting (You et al., 2023), etc. Hence, the limitations in functionality hinder the wide 046 applications of prior methods. 047

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 (Wu et al., 2024d) performs well on TID2013 dataset (Ponomarenko et al., 2015) (85.0%), but drops significantly to 50.7% when testing on BAPPS dataset (Zhang et al., 2018). A more comprehensive comparison on our newly created benchmark is given in Fig. 1, where it shows that previous works (Wu et al., 2024b;d) 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 (Wu

060

065 066

067

068

069

071

072

054



Figure 1: **Performance comparison**. Our model surpasses previous works including Q-Instruct (Wu et al., 2024b), Co-Instruct (Wu et al., 2024d), and the proprietary GPT-4V (OpenAI, 2023) across a broad range of tasks in both full-reference and non-reference settings. Traditional score-based IQA methods like LPIPS (Zhang et al., 2018) and MUSIQ (Ke et al., 2021) have no language abilities, and thus can only be used in *instant rating* task. Q-Instruct is only tested on single-image input tasks.

et al., 2024b) is limited; Co-Instruct (Wu et al., 2024d) directly utilizes GPT-4V (OpenAI, 2023),
which is not accurate in IQA tasks, to generate data; and the dataset scale in DepictQA (You et al., 2023) remains small. Additionally, these methods are constrained in their usage by resizing images to a fixed resolution (Wu et al., 2024b;d), while the image resolution is critical for quality assessment. 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 079 assessment tasks. We categorize these tasks into two types, as shown in Fig. 2. (a) Single-image assessment aims to evaluate the quality of a single image by identifying distortions (e.g., "blur" in 081 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 083 quality of two distorted images based on the clarity, colorfulness, and sharpness of presented contents. 084 For example, in Fig. 2b, despite reduced contrast, "Image A maintains more scene integrity", as 085 "Image B's serious noise level is more detrimental". We omit multi-image comparison since it is an easy extension of a pairwise one (Gu et al., 2020). Each type includes basic *brief* sub-tasks for 087 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. 088

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 (You et al., 2023). Moreover, our dataset is suitable for both full-reference and non-reference settings.

With EDQA-495K dataset, we then train a VLM model, named Enhanced Descriptive image Quality
 Assessment (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.

100 The performance of EDQA is evaluated in Fig. 1 and Sec. 5. In brief tasks, our model surpasses 101 general VLMs, IQA-specific VLMs, and score-based IQA methods by a large margin. For example, 102 we achieve 95.9% in non-reference comparison on TID2013 dataset, remarkably surpassing Co-103 Instruct (85.0%) and GPT-4V (88.1%). In detailed tasks, our model also excels, e.g., recording 104 74.9% in non-reference comparison reasoning, compared to 48.1% for Co-Instruct and 60.3% for 105 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-106 functional model could serve as a stepping stone towards a unified VLM-based IQA model. Although 107 not yet fully realized, our method showcases the potential of VLM-based IQA models.



Figure 2: **Task paradigm and qualitative results**. EDQA focuses on two 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 in Fig. A10, A11, A12, and A13.

# 123 2 RELATED WORKS

119

120

121

122

125 Score-based IOA methods. Traditional IOA 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 126 similarity score between a distorted image and a high-quality reference. Early works rely on human-127 designed metrics such as image information (Sheikh and Bovik, 2006), structural similarity (Wang 128 et al., 2004), phase congruency with gradient magnitude (Zhang et al., 2011), etc. The rapid 129 advancement of deep learning has also inspired learning-based IQA methods that measure image 130 quality through data-driven training. Pioneered by PieAPP (Prashnani et al., 2018) and LPIPS (Zhang 131 et al., 2018), data-driven approaches (Bosse et al., 2018; Cao et al., 2022; Ding et al., 2020; 2021; 132 Ghildyal and Liu, 2022; Yin et al., 2022; Zhou and Wang, 2022) have spurred innovations in IQA, 133 exhibiting high consistency with human judgments. (b) Non-reference methods directly regress 134 a quality score without a reference image. Initially, human-designed natural image statistics are 135 adopted (Ma et al., 2017; Mittal et al., 2012; 2013; Moorthy and Bovik, 2010; 2011; Saad et al., 2012; 136 Tang et al., 2011). Subsequently, deep-learning-based methods (Kang et al., 2014; Liu et al., 2017; Pan et al., 2018; Su et al., 2020; Sun et al., 2022; Zheng et al., 2021; Zhu et al., 2020) replace hand-137 crafted statistics by learning quality priors from extensive data. Recent works focus on enhancing 138 performance by introducing multi-scale features (Ke et al., 2021), CLIP pre-training (Wang et al., 139 2023), multi-dimension attention (Yang et al., 2022), continual learning (Zhang et al., 2022), multitask 140 learning (Zhang et al., 2023b), and so on. However, as discussed in You et al. (2023), score-based 141 IQA methods limit themselves in complex analyses and multi-aspect weighing of IQA, since the 142 information provided by a single score is far from sufficient. 143

Vision Language Models (VLMs) incorporate visual modality into large language models (Chiang 144 et al., 2023; Openai, 2023; Touvron et al., 2023a), aiming to leverage their emergent ability to achieve 145 general visual ability. These VLMs (Alayrac et al., 2022; Dai et al., 2023; Liu et al., 2023a; OpenAI, 146 2023; Wei et al., 2023; Ye et al., 2023a; Yin et al., 2023; Zhang et al., 2023a; 2024; Zhu et al., 2024a) 147 have demonstrated a general visual ability and can tackle a variety of multi-modality tasks, including 148 image captioning (Agrawal et al., 2019; Chen et al., 2015; Young et al., 2014), visual question 149 answering (Goyal et al., 2017; Liu et al., 2023b; Lu et al., 2022), document understanding (Masry 150 et al., 2022; Mathew et al., 2021; Singh et al., 2019), etc. Although proficient in these high-level 151 perception tasks, we demonstrate in Sec. 5 that general-purpose VLMs still struggle with IQA tasks.

152 **VLM-based IQA methods** aim to achieve better alignment with human perception leveraging 153 the power of VLMs (Wu et al., 2024e). Q-Bench (Wu et al., 2024a) establishes a comprehensive 154 benchmark for evaluating general-purpose VLMs in low-level perception tasks. Zhu et al. (2024b) evaluates various VLMs on the widely-adopted two-alternative forced choice (2AFC) task. Q-156 Instruct (Wu et al., 2024b) enhances the low-level perception ability of VLMs by introducing a 157 large-scale dataset. Q-Align (Wu et al., 2024c) employs discrete text-defined levels for more accurate 158 quality score regression. Co-Instruct (Wu et al., 2024d) concentrates on the quality comparison among multiple images. DepictQA (You et al., 2023) performs quality description, quality comparison, 159 and comparison reasoning in the full-reference setting. Nonetheless, as highlighted in Sec. 1, these methods focus only on specific aspects of IQA tasks, diverging from the original intents of VLMs' 161 universality and practical usage requirements, and their performance remains sub-optimal.

162 Table 1: 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

171

172

173

174

175

176

177

178

179

181

182

183

184

185

186

187

188

189

190

#### TASK PARADIGM AND DATASET CONSTRUCTION 3

#### 169 3.1 TASK PARADIGM 170

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 (Gu et al., 2020).

191 Compared with previous works, our design unifies various tasks, response types, and reference 192 settings into a multi-functional paradigm. In contrast, Q-Instruct (Wu et al., 2024b) focuses on 193 non-reference single-image assessment, Co-Instruct (Wu et al., 2024d) targets comparison among 194 multiple images in the non-reference setting, and DepictQA (You et al., 2023) primarily addresses 195 the full-reference setting. Although one can achieve unified IQA by combining these task-specific 196 IQA models, it is impractical due to the significant increase in network parameters, considering that 197 current VLMs are already quite large.

#### 199 3.2 DISTORTION LIBRARY 200

Existing IQA datasets (e.g., BAPPS (Zhang et al., 2018), PieAPP (Prashnani et al., 2018)) usually 201 introduce distortions (e.g., noise, blur) into high-quality reference images to create distorted images 202 for evaluation. However, these datasets do not publicly release the distortion information of each 203 image, and their distortions only cover limited scenes. Therefore, we aim to develop a comprehensive 204 large-scale distortion library. 205

206 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 207 "Gaussian blur", "motion blur", "lens blur", etc. In total, there are 35 sub-categories. For each sub-208 category, there are 5 severity levels: "slight", "moderate", "obvious", "serious", and "catastrophic". 209 A summary is illustrated in Tab. 1. Considering the need to assess high-quality images as well, we 210 retain the original image without any distortions in 5% proportion. See details in Appendix B.1. 211

212 Multi-distortion setups. In practical usage, multiple distortions may occur simultaneously on 213 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" 214 weakens "darken", "blur" weakens "over-sharpen". Second, certain distortions exhibit similar visual 215 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 Fig. A3. Hence, we limit the distortion number to two and
 manually review all possible combinations to exclude contradictory or similar combinations. See
 details of multi-distortion setting in Tab. A1 and Appendix B.1.

220 221

222

3.3 DATASET CONSTRUCTION

High-quality and large-scale datasets are crucial for training VLMs. Following Liu et al. (2023a);
Yin et al. (2023), 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.

228 Image collection. Typical IQA datasets involve two types of images: high-quality reference images 229 and distorted images to be evaluated. Generating distorted images is easy given our comprehensive 230 distortion library introduced in Sec. 3.2. Existing studies often collect a large number of distorted 231 images from a small number of references (Gu et al., 2020; Ponomarenko et al., 2015). However, the semantic richness of images is also crucial for VLM training. Therefore, we primarily source 232 reference images from the KADIS-700K dataset (Lin et al., 2020), which offers 140K pristine 233 reference images from diverse natural and daily scenes. We also leverage other IQA datasets for their 234 convenience to generate responses (details are below). 235

Question collection. Humans often express similar questions using different sentences, necessitating
 model robustness to various user questions. For each task, we initially prompt GPT-4 (Openai, 2023)
 to generate 50 candidate questions. Subsequently, we manually eliminate ambiguous and repetitive
 ones and correct inaccurate ones, creating a question set of 20 questions (see Appendix B.2). These
 questions are randomly sampled during training and testing to form the data pair.

241 **Response collection**. We employ two response types as shown in Fig. 3. The first comprises brief 242 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 243 enhance the model's advanced reasoning abilities. Existing methods to collect detailed responses 244 mainly rely on human annotation (Wu et al., 2024b; You et al., 2023) and GPT-4V generation (Wu 245 et al., 2024d). However, human annotation can be biased and vary in quality particularly when 246 annotators are untrained or tired (You et al., 2023). Also, GPT-4V is not fully reliable since its IQA 247 performance is still unsatisfactory as evidenced in Sec. 5. 248

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 (*e.g.*, comparison results). While GPT-4V excels at identifying contents and analyzing impacts, it struggles with distortion identification and quality comparison, which will be shown in Sec. 5. 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 to complete the
response. For streamlined evaluation, we randomly select 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. Here prior studies (Wu et al., 2024b;d) primarily focus on low-level properties, while we consider how these low-level distortions influence the display of high-level contents.

*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



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.

of 20 templates to convert the comparison results into textural responses, and append the short
answer prompt for the convenience of evaluation. We select three IQA datasets for training, including
BAPPS (Zhang et al., 2018), KADID-10K (Lin et al., 2019), and PIPAL (Gu et al., 2020), 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. 4 and Tab. 8, 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.

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", "saturate weaken", "saturate strengthen", "quantize", and "over-sharpen".

308 Dataset statistics. The dataset statistics are
309 illustrated in Tab. 2 (more in Appendix B.2).
310 All tasks, except *instant rating*, are displayed in
311 the single-distortion / multi-distortion format.
312 Our training set contains 439,676 brief samples
313 and 55,577 detailed samples. For *instant rating*,
314 the training set includes BAPPS, KADID, and

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

PIPAL, while the validation set consists of BAPPS, KADID, PIPAL, TID2013 (Ponomarenko et al., 2015), LIVE-MD (Jayaraman et al., 2012), and MDID2013 (Sun et al., 2017). To ensure 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.

319

4 MODEL DESIGN

320 321

We primarily follow LLaVA-1.6 (Liu et al., 2024) and mPLUG-Owl2 (Ye et al., 2023b) to construct our model, which is composed of a vision encoder, vision abstractor, and LLM. The vision encoder is a frozen CLIP pre-trained ViT-L/14 (Radford et al., 2021). The vision abstractor is a four-layer transformer, reducing the number of vision tokens to 64 and mapping the vision tokens to textual
space. The base LLM is Vicuna-v1.5-7B (Chiang et al., 2023). Following You et al. (2023), we also
adopt the unique tag technique, *i.e.*, using specific tags for various types of images, to ensure that the
language model can distinguish different input images. To increase robustness, an external high-level
perception dataset (detailed description dataset in Yin et al. (2023)) is incorporated during training as
a regularization, similar to Wu et al. (2024b); You et al. (2023). See model details in Appendix C. Our
experiments in Tab. A9 also show that model architecture has little influence on model performance.

Retaining resolution in training. Although previous VLM-based IQA models typically resize all input images to a fixed resolution (Wu et al., 2024b;d), 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 (Radford et al., 2021) to accommodate varying image resolutions. Ablation studies detailed in Sec. 5.4 demonstrate our model's capability to assess quality variations attributable to resolution, even without explicitly training on such tasks.

338 **Confidence estimation.** In many applications, it is important to know a confidence score that 339 indicates when the model is uncertain of its response. Here we use the confidence scores of some 340 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 341 reasoning tasks, which feature diverse and non-structured responses, we utilize semantic change 342 testing (Duan et al., 2023) to identify the top 20 tokens with the highest importance scores as key 343 tokens. In semantic change testing, we employ all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) as 344 the similarity model, due to its high processing speed (14K sentences per second). The predicted 345 likelihood of key tokens is averaged as the confidence score. Fig. 6 and Fig. A6 verify that confidence 346 and model performance are highly correlated. 347

348 349

350

- 5 EXPERIMENTS
- 351 5.1 METRICS AND BASELINES

352 353 Accuracy, SRCC, and PLCC. The accuracy metric is utilized for *distortion identification* and *instant* rating tasks. VLMs usually produce diverse textual outputs, and we transform them into brief results 354 for accuracy calculation. Specifically, we prompt our EDQA with "Answer the question using a 355 single word or phrase" to encourage direct output of brief responses. For baseline models, we include 356 all potential answers in the prompt and instruct the model to identify the most accurate one. We 357 emphasize that our key motivation is to generate descriptive language rather than quality scores. 358 However, our approach can produce quality scores using pair-wise comparison if required. The 359 quality scores are assessed using Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear 360 Correlation Coefficient (PLCC). The results of quality score are given in Appendix D.1. 361

**GPT-4 score, BLEU, and ROUGE-L**. We employ the GPT-4 score to evaluate *assessment reasoning* and *comparison reasoning* tasks, following Liu et al. (2023a). 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 See et al. (2017); Vaswani (2017).

369 **Baselines**. We categorize our baseline methods into general-purpose VLMs and IQA-specific VLMs. 370 For general VLMs, we include mPLUG-Owl2 (Ye et al., 2023b) (based on LLaMA-2-7B (Touvron 371 et al., 2023b)), LLaVA-1.6 (Liu et al., 2024) (based on Vicuna-v1.5-7B (Chiang et al., 2023)), and 372 the proprietary GPT-4V (OpenAI, 2023). IQA-specific VLMs are represented by Q-Instruct (Wu 373 et al., 2024b), Co-Instruct (Wu et al., 2024d), and CLIP-like LIQE Zhang et al. (2023b). Note that 374 Q-Instruct only supports single-image inputs, thus we only test it on non-reference single-image 375 assessment tasks. Additionally, we compare traditional score-based IQA methods including fullreference ones (PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 376 2020)) and non-reference ones (NIQE (Mittal et al., 2013), ClipIQA (Wang et al., 2023), MUSIQ (Ke 377 et al., 2021), MANIQA (Yang et al., 2022)) in instant rating task and score regression experiments.

384 385 386

387

388

397

Table 3: Distortion identification results under both single-distortion and multi-distortion cases. The
accuracy metric is reported in the full-reference / non-reference settings. EDQA greatly outperforms
all baselines and maintains its high accuracy in out-of-distribution (OOD) setting.

		General VLM		IQA-specific VLM				
	mPLUG-0	Owl2 LLaVA-1.	6 GPT-4V	Q-Instruct	Co-Instruct	LIQE	EDQA	EDQA (OOD)
Single-dist Multi-dist.	.   10.1 / 11 10.8 / 10	1.6 14.0 / 15.3 0.7 12.0 / 12.	3 46.7 / 45.2 1 41.5 / 39.8	2 - / 15.5 3 - / 23.9	27.2 / 34.4 30.2 / 33.3	- / 33.1 - / 31.4	97.7 / 94.1 91.3 / 89.3	82.1 / 73.2 76.6 / 77.2

Table 4: **Instant rating results** on multiple benchmarks in the full-reference / non-reference setting with the accuracy metric. 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.

Me	Methods		$KADID^{\texttt{test}}$	PIPAL <sup>test</sup>	TID2013	LIVE-MD	MDID2013	Mean
Full-refer. Score-based IQA	PSNR SSIM LPIPS DISTS	68.9 / 69.7 / 79.4 / 79.7 /	78.7 / 77.1 / 79.7 / 85.8 /	80.9 / 82.6 / 84.2 / 84.6 /	85.0 / 78.7 / 86.6 / 87.0 /	89.7 / 88.1 / 91.3 / <b>93.1</b> /	78.0 / 76.8 / 85.4 / 88.5 /	80.2 / 78.8 / 84.4 / 86.5 /
Non-refer. Score-based IQA	NIQE ClipIQA MUSIQ MANIQA	/ 49.9 / 59.7 / 59.2 / 54.9	/ 66.9 / 75.8 / 76.1 / 68.4	/ 59.7 / 72.6 / 77.8 / 79.2	/ 65.0 / 85.8 / 80.1 / 77.3	/ 86.9 / 65.8 / 87.2 / 75.4	/ 82.2 / 47.0 / 81.1 / 63.5	/ 68.4 / 67.8 / 76.9 / 69.8
General VLM	mPLUG-Owl2 LLaVA-1.6 GPT-4V	50.1 / 50.1 54.1 / 56.2 70.3 / 63.2	50.6 / 50.8 50.4 / 51.9 83.2 / 81.5	49.6 / 49.6 52.0 / 52.6 78.5 / 78.2	48.6 / 48.5 54.2 / 57.0 84.4 / 88.1	49.9 / 50.1 54.4 / 56.5 79.6 / 72.7	50.6 / 50.5 54.3 / 53.1 70.6 / 67.6	49.9 / 49.9 53.2 / 54.6 77.8 / 75.2
IQA-specific VLM	Q-Instruct Co-Instruct EDQA (Ours)	- / 41.6 49.8 / 50.7 <b>84.7</b> / <b>82.4</b>	- / 81.7 52.0 / 82.4 <b>93.6</b> / <b>93.1</b>	- / 74.6 50.6 / 72.5 <b>90.5</b> / <b>90.0</b>	- / 88.8 59.3 / 85.0 <b>96.9</b> / <b>96.4</b>	- / 73.1 50.0 / 70.3 92.1 / <b>91.8</b>	- / 48.5 50.0 / 58.0 <b>90.0 / 89.6</b>	- / 68.1 52.0 / 69.8 <b>91.3 / 90.6</b>

### 5.2 RESULTS ON BENCHMARKS

403 Quantitative results of distortion identification are shown in Tab. 3. First, the performance of 404 Co-Instruct is stably superior in the non-reference setting compared to the full-reference setting, 405 attributed to its training without reference. Second, the performance of open-source general VLMs, 406 including mPLUG-Owl2 and LLaVA-1.6, is still limited, but the proprietary GPT-4V (OpenAI, 407 2023) outperforms other general-purpose VLMs and exceeds prior specialized IQA VLMs. Third, 408 EDQA significantly surpasses all baseline methods, demonstrating our model's efficacy. Finally, we 409 evaluate our model in an out-of-distribution (OOD) setting. Specifically, for a particular category 410 of distortion (e.g., noise), we use some sub-categories (e.g., Gaussian noise) during training, and 411 different sub-categories (e.g., impulse noise) for evaluation. Results in the last column of Tab. 3 show that our method maintains high accuracy even under such an OOD setting. 412

413 Quantitative results of instant rating are demonstrated in Tab. 4. First, in the full-reference 414 context, traditional score-based methods, even the simplest PSNR, outperform all general VLMs 415 including GPT-4V and prior IQA-specific VLMs, indicating the inadequacy of existing VLMs in 416 full-reference IQA tasks. Second, conversely, in the non-reference scenario, GPT-4V and Co-Instruct 417 excel beyond most score-based approaches, except MUSIQ. Third, Co-Instruct is trained on multiimage comparison tasks without reference, and thus its performance in full-reference setting drops 418 by quite a large margin. This further demonstrates the necessity of unifying full-reference and 419 non-reference settings. Finally, EDQA demonstrates superior performance across both settings by a 420 large margin, showcasing its substantial advantage. 421

422 Quantitative results of assessment reasoning and comparison reasoning are shown in Tab. 5 and Tab. 6. First, the performance of the VLM-specific models significantly declines on tasks outside 423 their defined scopes. For instance, Co-Instruct's performance is unsatisfactory on full-reference 424 tasks. Second, GPT-4V shows robust reasoning abilities, stably outperforming prior IQA-specific 425 VLMs. Third, EDQA surpasses GPT-4V, especially in the non-reference setting, affirming its superior 426 reasoning abilities. Finally, EDQA achieves relatively good GPT-4 score and ROUGE-L, indicative 427 of the overall semantic accuracy, but a low BLEU score (yet remains much higher than GPT-4V), 428 which reflects word-level consistency. This suggests that while our predicted answers do not precisely 429 duplicate the ground truths word-for-word, they preserve similar meanings with diverse expressions. 430

**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 Appendix D.4 and Fig. A10, A11, A12, A13, A14.

 

 Table 5: Assessment reasoning and comparison reasoning results under both single-distortion and multi-distortion cases. GPT-4 score metric is reported in the full-reference / non-reference setting.
 EDC

 Methods
 Assessment Reasoning
 Comparison Reasoning

Methods	Assessment	Reasoning	Comparison Reasoning		
	Single-distortion	Multi-distortion	Single-distortion	Multi-distortion	
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	



Figure 4: User study.

Table 6: Assessment reasoning and comparison reasoning results with classic metrics (BLEU and ROUGE-L) in the full-reference / non-reference setting.

			Assessmen	t Reasoning		Comparison Reasoning				
	Methods	Single-d	listortion	Multi-d	istortion	Single-d	listortion	Multi-d	istortion	
		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	
E	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	

**Quality score regression**. Thought *our key motivation is to generate descriptive language rather than quality scores*, our approach can produce quality scores if required. We transform the score regression problem into instant rating tasks, and calculate the win rate of one image against others (selected by round robin for a small number, random sampling for a large number) as the quality score. The details and results are give in Appendix D.1.

458 5.3 REAL-WORLD APPLICATIONS

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. 5 and Fig. A14 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. 5d, 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. 4, revealing that our approach stably outperforms baseline methods in aligning human perception. 

Comparison on model-processed images. To
develop image restoration models, one often needs
to compare the restoration quality of different models. To simulate this, we consider five distortions
including "defocus blur", "motion blur", "noise",

Table 7: Results on model-processed images.

	NIQE	ClipIQA	MUSIQ	ManIQA	GPT-4V	EDQA
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

"JPEG compression", and "low resolution". For each distortion, three to four candidate models are used to process the distorted images. We manually rank the restored results, assigning "1" to the best restoration, "2" to the second best, etc. Different IQA methods are adopted to compare these restored images pairwise and find the best restoration. The average rank of the found best restoration and the accuracy of the paired comparison are reported in Tab. 7. 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. See details in Tab. A13 and Fig. A8. 

## 483 5.4 ABLATION STUDIES

**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



Figure 5: Qualitative results on assessing web-downloaded images. More results in Fig. A14.

Table 8: Our assistant model Table 9: Retaining resolution is surpasses GPT-4V greatly in important to identify the images instant rating task. The metric with better aspect ratio or higher is accuracy in the full-reference resolution. / non-reference setting.

	GPT-4V	Our Assistant	Retain	esolution? Inference	H↔W	$0.8 \times$	$0.9 \times$
TID2013	84.4 / 88.1	94.9 / 94.6	X	X	73.0	91.7	77.2
LIVE-MD	79.6 / 72.7	93.1 / 92.8	1	X	85.6	99.0	94.8
MDID2013	70.6 / 67.6	90.1 / 89.8	1	1	98.8	99.3	96.8



Figure 6: Confidence and performance are highly correlated.

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. 8 affirm the superiority of the assistant model.

**Retaining resolution**. In Tab. 9, we study the effects of retaining resolution. We sample 1,000 514 high-quality images with an aspect ratio greater than 4:3. These images are either resized by 515 swapping their height and width ( $H\leftrightarrow W$ ) or down-sampled by a scale factor of 0.8 or 0.9. The model 516 needs to compare the original and resized images to determine the better one. The alternative of 517 retaining resolution is resizing the two images to a larger resolution, which can maintain the quality 518 difference between the original and resized images (v.s., resizing to a smaller resolution results in two 519 nearly the same images). The results in Tab. 9 prove that retaining resolution is crucial for identifying 520 images with better aspect ratio or higher resolution. More results in Appendix D.2 and Tab. A8.

**Confidence**. We examine the correlation between model performance and estimated confidence 522 in Fig. 6. For *distortion identification* and *instant rating* tasks, across both full-reference and nonreference settings, our model demonstrates improved performance as the confidence interval increases. This validates the effectiveness of our confidence estimation. Details in Appendix D.2 and Fig. A6.

Model architecture is studied in Appendix D.2 and Tab. A9, showing little influence on performance.

526 527 528

529

521

523

524

525

500

501

502

503

504

505

506

507

508

509

510 511

512

513

#### 6 **CONCLUSIONS AND LIMITATIONS**

530 We introduce EDQA, a VLM-based IQA model, empowered by a new multi-functional task paradigm, 531 dataset enrichment, and training technique, surpassing baseline methods in both benchmarks and two real-world applications, showing the potential of descriptive quality assessment. 532

**Limitations.** First, the fine-grained abilities requiring more high-level perception skills are still 534 unsatisfactory. For example, in Fig. 5c, though identifying noise and pixelation successfully, our 535 model fails to point out that they are respectively located in the left and right parts. One possible 536 solution is to take the segmentation model to add various distortions to different regions. Second, for 537 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 introduced 538 to increase linguistic diversity during dataset construction. Third, whether our assessment can be 539 used as feedback to improve the quality of generation or restoration models is still under-explored.

# 540 REFERENCES

552

553

554

555

558

561

564

565

566

567

569

570

571 572

573

578

579

580

- H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson.
   Nocaps: Novel object captioning at scale. In *ICCV*, 2019.
- J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds,
   et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.
- 546
  547
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
  548
- 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 *CVPR*, 2022.
  - X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
  - X. Chen, Z. Li, Y. Pu, Y. Liu, J. Zhou, Y. Qiao, and C. Dong. A comparative study of image restoration networks for general backbone network design. *arXiv preprint arXiv:2310.11881*, 2023a.
- X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023b.
- W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, 2023. URL https://vicuna.lmsys.org.
- W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
  - K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020.
  - K. Ding, Y. Liu, X. Zou, S. Wang, and K. Ma. Locally adaptive structure and texture similarity for image quality assessment. In *ACM MM*, 2021.
  - J. Duan, H. Cheng, S. Wang, C. Wang, A. Zavalny, R. Xu, B. Kailkhura, and K. Xu. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*, 2023.
  - Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020.
- A. Ghildyal and F. Liu. Shift-tolerant perceptual similarity metric. In *ECCV*, 2022.
- 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 *CVPR*, 2017.
  - 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 ECCV, 2020.
- V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 2020.
- E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.
- D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. Objective quality assessment of multiply distorted images. In *Conference record of the forty sixth Asilomar conference on signals, systems and computers (ASILOMAR)*, 2012.
- J. Jiang, K. Zhang, and R. Timofte. Towards flexible blind jpeg artifacts removal. In *ICCV*, 2021.
- L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, 2014.
- 593 J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *CVPR*, 2021.

594 595 596	T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , 2018.
597 598 599	J. Lee, H. Son, J. Rim, S. Cho, and S. Lee. Iterative filter adaptive network for single image defocus deblurring. In <i>CVPR</i> , 2021.
600 601	J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. SwinIR: Image restoration using swin transformer. In <i>ICCVW</i> , 2021.
602 603 604	H. Lin, V. Hosu, and D. Saupe. KADID-10k: A large-scale artificially distorted iqa database. In <i>International Conference on Quality of Multimedia Experience (QoMEX)</i> , 2019.
605 606	H. Lin, V. Hosu, and D. Saupe. DeepFL-IQA: Weak supervision for deep iqa feature learning. <i>arXiv preprint arXiv:2001.08113</i> , 2020.
607 608	H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In NeurIPS, 2023a.
609 610	H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.
611 612 613	X. Liu, J. van de Weijer, and A. D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In <i>ICCV</i> , 2017.
614 615	Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? <i>arXiv preprint arXiv:2307.06281</i> , 2023b.
616 617	P. Lu, S. Mishra, T. Xia, L. Qiu, KW. Chang, SC. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In <i>NeurIPS</i> , 2022.
619 620	C. Ma, CY. Yang, X. Yang, and MH. Yang. Learning a no-reference quality metric for single-image super-resolution. <i>Computer Vision and Image Understanding</i> , 2017.
621 622	A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In <i>ACL Findings</i> , 2022.
623 624	M. Mathew, D. Karatzas, and C. Jawahar. DocVQA: A dataset for vqa on document images. In WACV, 2021.
625 626	A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. <i>IEEE TIP</i> , 2012.
627 628 629	A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. <i>IEEE Sign. Process. Letters</i> , 2013.
630 631	A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. <i>IEEE Sign. Process. Letters</i> , 2010.
632 633	A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. <i>IEEE TIP</i> , 2011.
635	Openai. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
636 637 638	OpenAI. GPT-4V(ision) system card, 2023. URL https://openai.com/research/ gpt-4v-system-card.
639 640	D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang. Blind predicting similar quality map for image quality assessment. In <i>CVPR</i> , 2018.
641 642 643	N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. <i>Signal processing: Image</i> <i>communication</i> , 2015.
644 645 646	E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In CVPR, 2018.

647 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

648	N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings
649	of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International
650	Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
651	
652	L. Ruan, B. Chen, J. Li, and M. Lam. Learning to deblur using light field generated and real defocus images. In
653	CVPR, 2022.
654	M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach
655	in the dct domain. <i>IEEE TIP</i> , 2012.
656	
050	A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In ACL,
007	2017.
658	H. R. Sheikh and A. C. Bovik. Image information and visual quality. <i>IEEE TIP</i> , 2006.
659	
660	A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards VQA
661	models that can read. In CVPR, 2019.
662	S Su O Yan Y Zhu C Zhang X Ge I Sun and Y Zhang Blindly assess image quality in the wild guided
663	by a self-adaptive hyper network. In <i>CVPR</i> , 2020.
664	- J
665	S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen. GraphIQA: Learning distortion graph representations for blind
666	image quality assessment. IEEE TMM, 2022.
667	W Sun E Zhou and O Liao Mdid: A multiply distorted image database for image quality assessment $PR$
669	2017
000	2017.
669	H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In CVPR, 2011.
670	U Terrere T Level C Level V Mertinet M A Lechene T Level D Desilier N Court D Herebry
671	H. IOUVION, I. LAVIII, G. IZACAIG, A. Martinel, MA. LaCIAUX, I. LACIOIX, B. KOZIETE, N. GOYAI, E. HAMOFO, E. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13071
672	2023a
673	20204.
674	H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava,
675	S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288,
676	2023b.
677	Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Boyik, and Y. Li, MAXIM: Multi-axis mln for image
678	processing. In CVPR. 2022.
679	1 ····· 8· ··· 9
620	A. Vaswani. Attention is all you need. In <i>NeurIPS</i> , 2017.
601	I Wang K C Chan and C C Loy Evploring clip for assessing the look and feel of images. In AAAL 2023
001	J. Wang, R. C. Chan, and C. C. Loy. Exploring cup for assessing the look and feel of images. In AAA1, 2025.
682	Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to
683	structural similarity. <i>IEEE TIP</i> , 2004.
684	H Wai I Kang I Chan I Theo 7 Co I Yang I Sun C Han and Y Thang Very Sealing up the vision
685	n. wel, L. Kolig, J. Chen, L. Zhao, Z. Ge, J. Talig, J. Sull, C. Han, and A. Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. <i>arXiv preprint arXiv:2312.06109</i> , 2023
686	vocability for farge vision-fanguage models. <i>urxiv preprint arxiv.2512.00107</i> , 2025.
687	H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai, et al. Q-bench: A
688	benchmark for general-purpose foundation models on low-level vision. In ICLR, 2024a.
689	H Wu Z Zhang E Zhang C Chan I Ling A Wang V Vu C Li I Hay C Zhai et al. O Instructu
690	II. WU, Z. Zhang, E. Zhang, C. Chen, L. Llao, A. Wang, K. AU, C. Li, J. HOU, G. Zhai, et al. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In CVPR 2024b
691	Improving four-level visual domines for math-modality foundation models. In $CVTR$ , $20240$ .
692	H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, et al. Q-Align:
603	Teaching lmms for visual scoring via discrete text-defined levels. In ICML, 2024c.
60/	H Wu H 7hu 7 7hang F 7hang C Chan I Liza C Li A Wang W Sun O Van stal Towards
605	open-ended visual quality comparison, arXiv preprint arXiv:2402 16641 2024d
092	
096	T. Wu, K. Ma, J. Liang, Y. Yang, and L. Zhang. A comprehensive study of multimodal large language models
697	for image quality assessment. arXiv preprint arXiv:2403.10854, 2024e.
698	S Yang T Wu S Shi S Lao Y Gong M Cao I Wang and Y Yang Manigas Multi-dimension attention
699	network for no-reference image quality assessment. In CVPR. 2022.
700	
701	Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. <i>arXiv preprint arXiv:2304.14178</i> , 2023a.

702 703	. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolution multi-modal large language model with modality collaboration. <i>arXiv preprint arXiv:2311.04257</i> , 202	iizing 23b.
704 705 706	. Yin, W. Wang, Z. Yuan, C. Han, W. Ji, S. Sun, and C. Wang. Content-variant reference image quassessment via knowledge distillation. In AAAI, 2022.	uality
708 707 708	Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, et al. Lamm: Lang assisted multi-modal instruction-tuning dataset, framework, and benchmark. In <i>NeurIPS</i> , 2023.	uage-
709 710	You, Z. Li, J. Gu, Z. Yin, T. Xue, and C. Dong. Depicting beyond scores: Advancing image quality assess through multi-modal language models. <i>arXiv preprint arXiv:2312.08962</i> , 2023.	sment
711 712 713	Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New simi metrics for semantic inference over event descriptions. <i>Transactions of the Association for Computat</i> <i>Linguistics</i> , 2014.	ilarity tional
714 715	W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, MH. Yang, and L. Shao. Multi-stage progressive i restoration. In <i>CVPR</i> , 2021.	mage
716 717	W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and MH. Yang. Restormer: Efficient transform high-resolution image restoration. In <i>CVPR</i> , 2022.	er for
718 719	Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assess <i>IEEE TIP</i> , 2011.	ment.
720 721 722 723	Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, W. Z H. Yan, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang. Internlm-xcom A vision-language large model for advanced text-image comprehension and composition. <i>arXiv pre arXiv:2309.15112</i> , 2023a.	hang, poser: print
724 725	Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep featu a perceptual metric. In <i>CVPR</i> , 2018.	res as
726 727 728	Zhang, J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao. LLaMA-Adapter: Efficient fine-tuni large language models with zero-initialized attention. In <i>ICLR</i> , 2024.	ng of
729 730	<sup>'</sup> . Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma. Continual learning for blind image quality assess <i>IEEE TPAMI</i> , 2022.	ment.
731 732	<sup>'</sup> . Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma. Blind image quality assessment via vision-lang correspondence: A multitask learning perspective. In <i>CVPR</i> , 2023b.	guage
733 734	. Zheng, J. Fu, Y. Zeng, ZJ. Zha, and J. Luo. Learning conditional knowledge distillation for degr reference image quality assessment. <i>ICCV</i> , 2021.	aded-
735 736	<sup>7</sup> . Zhou and Z. Wang. Quality assessment of image super-resolution: Balancing deterministic and stati fidelity. In <i>ACM MM</i> , 2022.	stical
737 738 739	. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. MiniGPT-4: Enhancing vision-language understanding advanced large language models. In <i>ICLR</i> , 2024a.	g with
740 741	. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. MetaIQA: deep meta-learning for no-reference image quassessment. In CVPR, 2020.	uality
742 743 744	. Zhu, X. Sui, B. Chen, X. Liu, P. Chen, Y. Fang, and S. Wang. 2afc prompting of large multimodal mode image quality assessment. <i>arXiv preprint arXiv:2402.01162</i> , 2024b.	ls for
745 746	PPENDIX	
747 748	OVERVIEW	
749 750 751 752	nis Appendix is structured as follows. Dataset details are described in Appendix B, followed b stails of model, training and inference in Appendix C. More ablation studies, qualitative results the details of real-world applications are presented in Appendix D.	y the
753 754	DATASET DETAILS	

B.1 DETAILS OF DISTORTION LIBRARY

756 As stated in Sec. 3.2, to facilitate the dataset construction, we design and implement a comprehensive distortion library. Our 758 distortion system contains 12 distortion super-categories in total, 759 with each category consisting of multiple sub-categories. For 760 instance, the "blur" category encompasses "Gaussian blur", "motion blur", "lens blur", etc. In total, there are 35 sub-categories. For 761 each sub-category, there are 5 severity levels: "slight", "moderate", 762 "obvious", "serious", and "catastrophic". In this section, we elaborate on our distortion implementations, including the principles, 764 formulas, and severity setup. We also provide one example for each 765 implementation in Fig. A2, with the reference image in Fig. A1. 766



Figure A1: Reference for distorted images in Fig. A2.

- Blur.
- · Gaussian blur. The distorted image is generated by convolving the reference image with a Gaussian blur kernel. We set the kernel size  $(s_k)$  to be a function of the standard deviation  $(\sigma_k)$ of the blur kernel:  $s_k = \text{round}(4 \times \sigma_k) + 1$ .
  - Motion blur. Linear motion blur is applied to the reference image using the linear filter, where  $(r, \sigma) \in [(5, 3), (10, 5), (15, 7), (15, 9), (20, 12)].$
- Filter the image using a Gaussian filter, then randomly jitter each • Glass blur. pixel in the image by x pixels, and repeat this process n iterations.  $[\sigma, x, n] \in$ [(0.7, 1, 1), (0.9, 2, 1), (1.2, 2, 2), (1.4, 3, 2), (1.6, 4, 2)].
- Lens blur. This distortion uses the circular average filter, where  $r \in [1, 2, 4, 6, 8]$ .
- Zoom blur. The image is gradually zoomed in and overlaid to calculate the average.
- Jitter blur. Each pixel is randomly displaced by a shift of randim (-p, p) pixels both in x and y dimensions, with a total of 5 displacements, where  $p \in [1, 2, 3, 4, 5]$ .

### Noise.

- Gaussian noise in RGB space. Additive Gaussian noise is applied to each of the RGB channels of an image, where  $\sigma \in [0.05, 0.1, 0.15, 0.2, 0.25]$ .
- Gaussian noise in YCrCb space. Similar to the Gaussian noise in RGB space, this distortion is implemented in YCbCr space, where  $(\sigma_l, \sigma_r, \sigma_b)$  $\in$ [(0.05, 1, 1), (0.06, 1.45, 1.45), (0.07, 1.9, 1.9), (0.08, 2.35, 2.35), (0.09, 2.8, 2.8)].
- Speckle Noise. Speckle Noise is also known as Multiplicative Gaussian noise, where  $\sigma \in$ [0.14, 0.21, 0.28, 0.35, 0.42].
- Spatially correlated noise. The reference image is first corrupted by an additive Gaussian noise, which results in each pixel being corrupted by an independent and identically distributed noise pattern. The resultant image is then filtered with an average filter of kernel size  $3 \times 3$ , correlating the intensity of each pixel with those of the neighboring pixels. More specifically, the distorted image is given by:

$$I_D(x, y, c) = \frac{1}{|N_n|} \sum_{i \in N_n} (I_R(x_i, y_i, c_i) + N(x_i, y_i, c_i)),$$
(A1)

where  $I_D$  is the distorted image,  $I_R$  is the reference image,  $N_n$  is the set of neighboring pixels, and  $N(x, y, c) \sim \mathcal{N}(0, \sigma_a^2)$ .

- Poisson noise. This distortion generates Poisson noise based on the image pixel values, where  $intervals \in [80, 60, 40, 25, 15].$
- Impulse noise. Impulse noise is also known as salt and pepper noise. The density of the noise:  $d \in [0.01, 0.03, 0.05, 0.07, 0.10].$

### Compression.

• JPEG. The distorted image is a JPEG-compressed version of the reference image, where the parameter in Pillow, quality  $q \in [25, 18, 12, 8, 5]$ .

15

769 770 771

772

773 774

775

776 777

778

779

780

781 782

783 784

785

786

787

788

789

790

791

792

793

794

800

801 802

803

804

805

806

808

809

 JPEG 2000. This distortion is an advanced compression widely used, where the Pillow's parameter quality q ∈ [29, 27.5, 26, 24.5, 23].

### Brightness.

- Brightness shift in HSV space. The RGB image is mapped to HSV, and then we enhance and reduce the brightness by V channel, where  $\sigma \in [0.1, 0.2, 0.3, 0.4, 0.5]$  for Brightening and  $\sigma \in [-0.1, -0.2, -0.3, -0.4, -0.5]$  for darkening.
- Brightness shift in RGB space. We enhance and reduce the brightness in all channels, where  $\sigma \in [0.1, 0.15, 0.2, 0.27, 0.35]$  for Brightening and  $\sigma \in [-0.1, -0.15, -0.2, -0.27, -0.35]$  for darkening.
- Gamma Brightness tuning in HSV space. The RGB image is mapped to HSV space and then we enhanced and reduce the brightness by V channel with a gamma function, where  $\gamma \in [0.7, 0.58, 0.47, 0.36, 0.25]$  for brightening and  $\gamma \in [1.5, 1.8, 2.2, 2.7, 3.5]$  for darkening.

### Contrast.

- Contrast tuning by scaling. Given an input image  $I_{in}$ , there is a corresponding  $I_{mean}$ , which is a gray image in which each element is the mean of  $I_{mean}$ . The distorted image  $I_D$  is generated as following:  $I_D = I_{mean} * (1.0 \alpha) + I_{in} * \alpha$ , where  $\alpha \in [0.75, 0.6, 0.45, 0.3, 0.2]$  for strengthening and  $\alpha \in [1.4, 1.7, 2.1, 2.6, 4.0]$  for weakening.
- Contrast tuning by stretching. Contrast changing is performed as follows:  $I_D(x, y, c) = 1/(1 + (\frac{\bar{I}_C}{I_R(x,y,c)+\epsilon})\alpha)$ , where  $I_D$  is the distorted image,  $I_R$  is the reference image, and  $\bar{I}_C$  is the mean intensity for channel c.  $\alpha \in [1.0, 0.9, 0.8, 0.6, 0.4]$  for weakening, and  $\alpha \in [2.0, 4.0, 6.0, 8.0, 10.0]$  for strengthening.

### Saturate.

- Saturate tuning in HSV space. The reference image is firstly mapped into HSV space and then the S channel is scaled, where the scale factor  $s \in [0.7, 0.55, 0.4, 0.2, 0.0]$  for weakening and  $s \in [3.0, 6.0, 12.0, 20.0, 64.0]$  for enhancement.
- Saturate tuning in YCbCr space. The reference image  $I_R$  is firstly mapped into YCbCr space and then the distorted image  $I_D$  is generated like the following formulation:

$$I_D(x, y, Cb) = 128 + (I_R(x, y, Cb) - 128) \times s,$$
(A2)

$$I_D(x, y, Cr) = 128 + (I_R(x, y, Cr) - 128) \times s,$$
(A3)

where  $s \in [0.6, 0.4, 0.2, 0.1, 0.0]$  donates the scale factor for weakening and  $s \in [2.0, 3.0, 5.0, 8.0, 16.0]$  for strengthening.

**Over-sharpen**. The reference image  $I_R$  is firstly processed by a Gaussian blur kernel to generated a blurred image  $I_{blur}$ . Then the original image is over-sharpened with cv2.addWeighted $(I_R, 1 + \alpha, I_{blur}, -\alpha, 0)$ , where  $\alpha \in [2, 2.8, 4, 6, 8]$ .

**Pixelate**. The reference image is firstly down-sampled in BOX mode, then up-sampled to the original resolution in NEAREST mode, where the down-sampling factor  $\sigma \in [0.5, 0.4, 0.3, 0.25, 0.2]$ .

### Quantize.

- Color quantization using histogram equalization. The color elements are divided into an equal histogram for quantization, where the number of classes  $c \in [24, 16, 8, 6, 4]$ .
  - Color quantization using histogram median. This distortion is implemented by the function PIL.Image.Quantize.MEDIANCUT, where the number of classes  $c \in [20, 15, 10, 6, 3]$ .
- Color quantization using OTSU method, which is implemented by existing function skimage.filters.threshold\_multiotsu to generate thresholds. The number of classes  $c \in [15, 11, 8, 5, 3]$ .
- Multi-distortion setups. As discussed in Sec. 3.2, multiple distortions may occur simultaneously on
   the same image in practical usage. First, we observe that humans can identify at most two distortions
   when three or more are applied, as in Fig. A3, thus we limit the number of applied distortions to two.



Figure A2: **Distortion examples** of our distortion design. We showcase one example for each distortion implementation. The reference image is depicted in Fig. A1.

907

Second, some distortions could weaken each other's presentation (*e.g.*, "brighten" weakens "darken",
"blur" weakens "over-sharpen"). Also, certain distortions show similar visual effects (*e.g.*, "pixelate"
looks similar to "blur"), making it hard to identify both if applied simultaneously. Hence, to exclude
contradictory or similar distortion combinations, we manually review all possible combinations. All
feasible distortion combinations used in our dataset are provided in Tab. A1.

915 Out-of-distribution setups. In Tab. 3, we evaluate our model in an out-of-distribution (OOD) setting.
916 Specifically, for a particular category of distortion (*e.g.*, noise), we use some sub-categories (*e.g.*,
917 Poisson noise) during training, and different sub-categories (*e.g.*, impulse noise) for evaluation. Here we provide a detailed split of training distortions and evaluation distortions in Tab. A2.

Table A1: Multi-distortion setting	where we show	v all feasible distortion	n combinations.
------------------------------------	---------------	---------------------------	-----------------

919		
920	First Distortion	All Possible Second Distortions
921	Blur	Brighten, Compression, Contrast Strengthen, Contrast Weaken, Darken, Noise, Quantize, Saturate Strengthen,
922	Drighton	Saturate weaken Plus Compression Noise Divelete Quentize
	Brighten	Blur, Compression, Noise, Pixelate, Quantize
923	Compression	Blur, Brighten, Contrast Strengthen, Contrast Weaken, Darken, Noise, Saturate Strengthen, Saturate Weaken
924	Contrast Strengthen	Blur, Compression, Noise, Pixelate, Quantize
925	Contrast Weaken	Blur, Compression, Noise, Pixelate, Quantize
010	Darken	Blur, Compression, Noise, Pixelate, Quantize
926		Blur, Brighten, Compression, Contrast Strengthen, Contrast Weaken, Darken, Over-sharpen, Pixelate, Saturate
927	Noise	Strengthen, Saturate Weaken
000	Over-sharpen	Brighten
920	-	Brighten Contrast Strengthen Contrast Weaken Darken Noise Over-sharpen Quantize Saturate Strengthen
929	Pixelate	Saturate Weaken
020	0	Brighten, Contrast Strengthen, Contrast Weaken, Darken, Noise, Over-sharpen, Pixelate, Saturate Strengthen,
930	Quantize	Saturate Weaken
931	Saturate Strengthen	Blur, Compression, Noise, Over-sharpen, Pixelate, Quantize
932	Saturate Weaken	Blur, Compression, Noise, Over-sharpen, Pixelate, Quantize



Figure A3: Humans usually identify at most two distortions (blue) when three (purple) are applied.

B.2 DETAILS OF DATASET CONSTRUCTION

**Template pool**. As stated in Sec. 3.3, for brief tasks, the questions and answers are templated and sampled from a pool. The questions of detailed tasks are also sampled from a pool. The question pools and answer pools (if possible) of *distortion identification, instant rating, assessment reasoning*, and *comparison reasoning* tasks are given in Tab. A14, Tab. A16, Tab. A15, and Tab. A17, respectively.

Statistics of the response length in our EDQA-495K dataset are detailed in Appendix B.2. We
provide statistics on both word count and string length. For the *instant rating* task, there is no
distinction between single-distortion and multi-distortion cases. We also depict the word length
distribution of detailed reasoning responses in Fig. A4.

Wordcloud map of our EDQA-495K dataset is given in Fig. A5. We manually exclude "Image A" and "Image B", since they are constant proper nouns across all texts. The most frequent words in our EDQA-495K dataset (*e.g.*, "overly high", "color quantization", "high contrast", "high saturation", and "detail") are all highly relevant to the low-level properties and the visual quality of images.

957 958

959

918

940

941

942 943

944

945

946

947

948

C DETAILS OF MODEL SETUPS

960 Model Architecture. EDQA primarily adopts the architecture from LLaVA-1.6 (Liu et al., 2024) 961 and mPLUG-Owl2 (Ye et al., 2023b), structured as follows. Specifically, the input images and 962 the question texts are first tokenized, then fused, finally processed by the Large Language Model 963 (LLM) for response generation. (1) Tokenizing input images and question texts. We use a frozen 964 CLIP pre-trained ViT-L/14 (Radford et al., 2021) as the image encoder to convert the input images 965 into visual tokens. The question texts are tokenized into textual tokens using the SentencePiece 966 tokenizer (Kudo and Richardson, 2018). To bridge the different embedding spaces of visual and 967 textual tokens, we implement a trainable image abstractor, which is a four-layer transformer network, 968 to map visual tokens into the textual space following Ye et al. (2023b). The abstractor can also 969 significantly 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) 970 Response generation using LLM. The fused tokens are fed into LLM to generate the final response. 971 Here we mainly conduct experiments with Vicuna-v1.5-7B. Despite their capabilities, pre-trained

973 974	Category	Training distortions	Validation distortions
975	Blur	Motion blur, Glass blur, Lens blur, Zoom blur	Gaussian blur, Jitter blur
76	Noise	Gaussian noise in YCrCb space, Speckle noise, Spatial correlated noise, Poisson noise	Gaussian noise in RGB space, Impulse noise
77	Compression	JPEG compression	JPEG2000 compression
78	Brighten	Shift brighten in HSV & RGB spaces, Gamma brighten in HSV space	Gamma brighten in RGB space
'n	Darken	Shift darken in HSV & RGB spaces, Gamma darken in HSV space	Gamma darken in RGB space
5	Contrast strengthen	Contrast strengthen by scaling	Contrast strengthen by stretching
0	Contrast weaken	Contrast weaken by scaling	Contrast weaken by stretching
81	Saturate strengthen	Saturate strengthen in HSV space	Saturate strengthen in YCrCb space
2	Saturate weaken	Saturate weaken in HSV space	Saturate weaken in YCrCb space
}	Quantization	Quantization by OTSU method, Quantization by histogram median	Quantization by histogram equalization

### Table A2: Setting of out-of-distribution (OOD) distortion identification.

Table A3: **Response length statistics** in EDQA-495K dataset, reported as word count / string length. For *instant rating* task, there is no distinction between single-distortion and multi-distortion cases.

	Distortion Identification	Assessment Reasoning	Instant Rating	Comparison Reasoning
Single-distortion	10.36 / 69.81	64.37 / 430.23	0 20 / 52 02	93.20 / 604.97
Multi-distortion	12.84 / 88.67	87.31 / 588.44	9.307 32.02	114.04 / 740.68

LLMs typically do not perform well on IQA tasks without adjustments. Therefore, we employ
LoRA (Hu et al., 2021), 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 Hu et al. (2021); Yin et al. (2023). This approach allows for targeted refinement
of the model's performance on IQA tasks without the need for extensive retraining.

997 **Model setup.** Since the CLIP pre-trained ViT-L/14 (Radford et al., 2021) encodes each  $14 \times 14$ 998 patch to a visual token, the resolution of the input image should be integer multiples of 14. Therefore, 999 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 (Radford et al., 2021), with 1000 each token having a channel of 1024. The vision abstractor can reduce the number of vision tokens 1001 to 64 and map the vision tokens to the hidden dimension of the LLM, which is 4096. Without the 1002 vision abstractor, the maximum resolution is limited to 672, constrained by computation resources 1003 (RTX A6000 GPUs). However, with the vision abstractor, we can process images with much larger 1004 resolutions (up to  $2500 \times 2500$ ). In our experiments, the maximum image resolution is  $1092 \times 1456$ , 1005 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. There are 32 attention layers in the LLM in total. In each attention layer, the 1008 projection weights of "query", "key", "value", and "output" are adjusted using two delta parameters 1009 with the shape of  $4096 \times 16$  and  $16 \times 4096$ , respectively.

**Training and inference setup.** In our experiments, we set the LoRA rank to 16. 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.

1015

1017

972

984

985

986 987

989 990 991

### 1016 C.1 COMPLEXITY AND EFFICIENCY

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.

Inference cost. The inference latency depends on the response length and it is tested on a single RTX A6000 GPU. For example, 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).



Figure A4: Word length distribution of detailed Figure A5: Wordcloud map of our introduced responses in our EDQA-495K dataset. EDQA-495K dataset.

Table A4: **Results of quality score regression** with SRCC / PLCC metrics in full-reference setting.

Methods	PIPAL	KADID <sup>test</sup>	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	<b>0.937</b> / 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 / <b>0.949</b>

### D MORE RESULTS

1049 1050 1051

1039

1040 1041

### D.1 QUALITY SCORE REGRESSION

1052

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 LLMbased 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.

1059 Quality score regression. The score regression results are evaluated on the PIPAL, KADID, TID2013, and CSIO 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 1061 score. Specifically, for an image A, we randomly sample comparison candidates, such as B, C, D, 1062 etc., which share the same content as A but have different distortions. Image A is then compared 1063 pairwise with each of its comparison candidates (B, C, D, etc.). In the full-reference setting, the 1064 reference image, Image A, and one compared candidate are input into our model for comparison. In the no-reference setting, only Image A and its one compared candidate are input for comparison. Finally, the win rate of Image A against all its compared candidates is calculated as its quality score. 1067 The comparison numbers per image for PIPAL, KADID, TID2013, and CSIQ datasets are 58, 62, 60, 1068 and 15, respectively. We show that the comparison number per image could be reduced significantly 1069 without large performance degradation in Tab. A7. The results of quality score regression are given 1070 in Tab. A4 and Tab. A5, proving that our method can generate accurate quality scores.

1071 Assessing in-the-wild images with different contents. Existing real-world IQA datasets like 1072 KonIQ Hosu et al. (2020) and SPAQ (Fang et al., 2020) contain real distorted images with various 1073 contents. To regress quality scores from such a dataset, our model needs to compare images with 1074 different contents though it is trained only to compare images with similar contents, as shown in Task 1075 2 of Fig. 2. The results in Tab. A6 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 1077 1078 our EDQA, *i.e.*, trained on KonIQ then evaluated on SPAQ, and vice versa. Our re-trained EDQA outperforms all baseline score-based methods trained on the same dataset. These results indicate that 1079 our method is capable of assessing in-the-wild images with different contents.

1081		• 0			
1082	Methods	PIPAL	KADID <sup>test</sup>	TID2013	CSIQ
1083	NIQE	0.300 / 0.367	0.430 / 0.499	0.315 / 0.413	0.660 / 0.747
1084	CLIPIQA	0.448 / 0.491	0.644 / 0.653	0.616 / 0.690	0.761 / 0.798
1085	MUSIQ	0.539 / 0.570	0.650 / 0.668	0.578 / 0.693	0.755 / 0.811
1086	MANIQA	0.3387 0.002	0.4827 0.327	0.4727 0.005	0.70170.714
1087	EDQA (Ours)	0.742 / 0.778	0.937 / 0.941	0.847 / 0.866	0.912 / 0.938

<sup>1080</sup> Table A5: **Results of quality score regression** with SRCC / PLCC metrics in non-reference setting.

Table A6: Results of quality score calculation on SPAQ and KonIQ datasets with SRCC / PLCC metrics. "LoRA" means the LLM is tuned using LoRA technique, where only 0.24% parameters (16M) of whole LLM (6.76B) are trainable. 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) 1	Results on S	PAQ dataset		
Methods	NIQE	CLIPIQA	MUSIQ	MANIQA	Q-Align-LoRA	EDQA-LoRA	EDQA-LoRA
Train Set	-	-	KonIQ	KonIQ	KonIQ	KonIQ	Original
SRCC	0.664	0.700	0.856	0.755	0.854	0.859	0.835
PLCC	0.679	0.722	0.859	0.765	0.855	0.861	0.841
			(b) I	Results on K	onIQ dataset		
Methods	NIQE	CLIPIQA	DBCNN	MUSIQ	Q-Align-LoRA	EDQA-LoRA	EDQA-LoRA
Train Set	-	-	SPAQ	SPAQ	SPAQ	SPAQ	Original
SRCC	0.530	0.685	0.731	0.753	0.782	0.787	0.717
PLCC	0.533	0.717	0.758	0.680	0.802	0.807	0.729

1105 1106

1094 1095

1107 **Influence of comparison numbers.** We calculate the win rate of one image over other compared 1108 images as the quality score. Here the compared images are selected by round robin for a small 1109 number, and random sampling for a large number. For the SPAQ dataset, the number of possible 1110 compared images is quite large, thus we adopt the random sampling strategy. The influence of comparison numbers is investigated in Tab. A7. It is shown that the comparison number could be 1111 reduced significantly without large performance degradation. In the most extreme cases (*i.e.*, the 1112 comparison number is 1 or 2), we use the estimated confidence as weights to calculate the win rate as 1113 quality score. Otherwise, the values of the win rate are too discrete (i.e., the values of the win rate can 1114 only be 0 or 1 when the comparison number is 1). The results of our EDQA are still reasonable in 1115 such extreme cases. Considering that the random sampling may bring large randomness or variance, 1116 we average the results with 5 random runs for small comparison numbers (*i.e.*, < 10). Although the 1117 comparison number is small and the sampling process is random, our method is still very stable with 1118 relatively small standard deviations in Tab. A7.

1119

# 1120 D.2 MORE ABLATION STUDIES

1122 Ablation study on retaining resolution. In Tab. 9 of the main paper, the effects of retaining 1123 resolution is studied. Specially, we randomly sample 1,000 high-quality images whose aspect ratios 1124 are greater than 4 : 3. These images are either resized by swapping their height and width (denoted as 1125  $H\leftrightarrow W$ ), 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 1126 the superior one. Note that in Tab. A8, the alternative method of retaining resolution is to resize both 1127 original image and resized image to a larger resolution, which can maintain the quality difference. In 1128 contrast, resizing both images to smaller resolution results in two nearly same images. 1129

The results are presented in Tab. A8. First, overall, 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

Table A7: Influence of comparison numbers per image on SPAQ dataset with SRCC and PLCC metrics. "(KONIQ)" means the model is trained on KONIQ dataset, which is also an in-the-wild IQA dataset. For small comparison numbers (< 10), we average the results with 5 random runs.</li>

37 38	Comparison Nur	mbers   100	50	25	10	5	2	1
39 40	EDQA (Original)	SRCC   0.835 PLCC   0.841	0.832 0.837	0.826 0.832	0.806 0.810	$\substack{0.731 \pm 0.006 \\ 0.735 \pm 0.006}$	$\substack{0.647 \pm 0.009 \\ 0.639 \pm 0.009}$	$\substack{0.577 \pm 0.015 \\ 0.537 \pm 0.015}$
	EDQA (KONIQ)	SRCC   0.859 PLCC   0.861	0.854 0.858	0.850 0.852	0.830 0.833	$\substack{0.756 \pm 0.006 \\ 0.757 \pm 0.007}$	$0.664{\pm}0.011$ $0.652{\pm}0.011$	$\substack{0.598 \pm 0.013 \\ 0.546 \pm 0.015}$

Table A8: **Retaining resolution** during both training and inference is important to identify images with better aspect ratio or higher resolution.

Retain R Training	esolution? Inference	H↔W	0.5  imes	$0.75 \times$	0.8  imes	$0.85 \times$	$0.9 \times$	$0.95 \times$
×	×	73.0	99.0	93.5	91.7	83.8	77.2	71.2
1	×	85.6	99.8	99.4	99.0	95.9	94.8	89.4
1	1	98.8	99.9	99.6	99.3	99.1	96.8	97.0

( $\leq$  99.0). 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.

Ablation study on confidence estimation. We further examine the correlation between model
performance and estimated confidence scores on a wider range of benchmarks. The results are
illustrated in Fig. A6. The performance of our model is consistently enhanced as the confidence
interval increases, validating the effectiveness of our confidence estimation.

1159Ablation study on model architecture. We compare two vision-text connectors (*i.e.*, vision abstractor1160v.s. projector) and three LLMs on distortion identification and instant rating tasks. The default vision-1161text connector and LLM in this ablation study is vision projector and Vicuna-v1.5-7b. The results in1162Tab. A9 show that the performance is similar. Considering that vision abstractor can greatly reduce1163the computational burden than projector, we select abstractor in the main paper. For example, for a1164448 × 448 image, projector generates 1024 tokens, while abstractor only outputs 64 tokens. Note1165that the amount of computation is proportional to the square of the number of tokens.

Relationships between the comparison reasoning and instant rating tasks are studied in Tab. A10.
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.

1170 Influence of input order on quality comparison. Recently, Zhu et al. (2024b) shows that some 1171 VLMs are sensitive to the input order of paired images when comparing the quality of two images. We test our model on the fine-grained dataset released by Zhu et al. (2024b) in Tab. A11. We follow 1172 Zhu et al. (2024b) to report the consistency / accuracy / correlation as metrics. Consistency means 1173 consistency in changing the order of input images. Our model achieves more than 0.90 consistency in 1174 all splits. Also, our comparison accuracy and score correlation are both much higher than Q-Instruct 1175 and GPT-4V. These results show that our model is robust to the order of input images. The statistics 1176 of our model's confidence are given in Tab. A12. The results show that the confidence of consistent 1177 prediction is much higher than inconsistent prediction, reflecting the self-evaluation ability of EDQA. 1178

1179

1143

1144

D.3 DETAILS OF REAL-WORLD APPLICATIONS

Details of quality comparison on model-processed images. 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. A13), apply them to a correspondingly degraded image, and then manually rank the resultant model-processed images. To find the image



Figure A7: An example of the model-restored image.



Figure A6: **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.

Table A9: **Model architecture** (*i.e.*, vision-text connectors and LLMs) has relatively little influence on model performance.

Types	Architectures	Distortion Id	lentification			Instant Rating		
21		Single-dist. ID	Multi-dist. ID	$BAPPS^{\texttt{test}}$	$\textbf{KADID}^{\texttt{test}}$	PIPAL <sup>test</sup>   TID2013	LIVE-MD	MDID2013
Vision-text connector	Projector Abstractor	97.9 / 94.7 97.7 / 94.1	90.5 / 89.5 91.3 / 89.3	82.7 / 81.4 84.7 / 82.4	92.7 / 92.4 93.6 / 93.1	89.2 / 88.8   96.2 / 95.9 90.5 / 90.0   96.9 / 96.4	92.1/91.9 92.1/91.8	89.1 / 88.4 90.0 / 89.6
LLM	Vicuna-v1.5-7b Vicuna-v0-7b LLaMA-2-7b	97.9 / 94.7 96.9 / 93.6 97.0 / 94.0	90.5 / 89.5 89.8 / 89.3 90.6 / 89.1	82.7 / 81.4 82.5 / 81.3 81.7 / 81.5	92.7 / 92.4 92.8 / 92.2 92.6 / 92.0	89.2 / 88.896.2 / 95.988.2 / 88.195.0 / 94.788.4 / 87.994.6 / 94.1	92.1/91.9 91.2/91.1 91.8/91.1	89.1 / 88.4 90.5 / 90.2 90.9 / 90.7

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.

1212 We have shown that our EDQA can consistently find the near-optimal candidate compared to GPT-4V 1213 and scored-based methods. A highlight is that our EDQA generalizes well on these out-of-distribution 1214 (OOD) model-processed images. For example, the image in Fig. A7 is restored from a noisy image. 1215 There is still remnant noise, which is somewhat strange. For such an OOD image, our EDQA correctly 1216 recognizes it to be inferior, but MANIQA, MUSIQ, and NIQE consider it as the best of the four 1217 candidates. We provide two qualitative results of detailed comparison reasoning on model-processed 1218 images in Fig. A8. One compares SwinIR and FBCNN in the JPEG compression artifact removal 1219 task, and the other compares MPRNet and MAXIM in the deblur task. Our model can generate a reasonable explanation for the comparison results. 1220

### 1222 D.4 MORE QUALITATIVE RESULTS

More qualitative results of *assessment reasoning*, *comparison reasoning*, and assessment on webdownloaded images are presented in Fig. A10, Fig. A11, Fig. A12, Fig. A13, and Fig. A14. EDQA could accurately identify distortions, analyze their impacts on the display of image contents, then weigh the advantages and disadvantages of different aspects, and finally draw a final conclusion (*e.g.*, overall quality, comparison results). In Fig. A9, we also present some qualitative results on assessing images with extremely severe distortions. Our model performs well in such extreme cases.

1230

1221

1223

1201 1202 1203

- 1231 1232
- 1233
- 1234
- 1235
- 1236
- 1237
- 1238
- 1239 1240
- 1241

			(a)	) Results	s on the	instant	rating	task					
	BAP	$PS^{test}$	KADID <sup>te</sup>	est PIP	$AL^{test}$	TID2	.013	LIVE-	MD	MDI	D201	3 M	ean
Only Rati Co-traini	ing   81.6 ng   <b>84.7</b>	/ 81.6 / <b>82.4</b>	92.4 / 92 <b>93.6</b> / <b>93</b>	.3 89.1 .1 90.5	1 / 89.0 5 / <b>90.0</b>	94.2 / 96.9 /	94.1 <b>96.4</b>	<b>92.9 / 9</b> 92.1 / 9	<b>92.7</b> 91.8	<b>92.1</b> 90.0	/ <b>91.</b> / 89.	7   90.4 6   <b>91.3</b>	/ 90.2 / <b>90.6</b>
			(b) Res	sults on t	the com	parison	reasor	ing tasl	k				
	G	PT-4 Sco	Single ore B	-distortio BLEU	on ROU	JGE-L	GP	Г-4 Scoi	Mı re	ulti-di BLI	storti EU	on ROU	JGE-L
Only Reas Co-train	oning 7 ing <b>7</b>	4.3 / 69. 7 <b>5.1</b> / <b>74.</b>	6 0.203 9 <b>0.20</b> 7	3 / 0.202 7 / <b>0.207</b>	0.465 0.466	/ 0.453 / <b>0.463</b>	70 71	.6 / <b>69.</b> 1 .7 / 68.7	1 0 7 0	.165 / <b>.176</b> /	0.16 0.17	5 0.414 2 0.420	/ 0.40 / <b>0.41</b>
Fable A11     with consist	: Quality stency / a	y <b>comp</b> a locuracy	arison ro	<b>esults</b> o ation me	on the fi etrics.	ne-gra	ined d	ataset	relea	sed b	y Zh	u et al.	(2024)
Datasets		Setting		Q	-Instrcu	ıt		GPT-4	١V		E	DQA (C	urs)
CSIQ CSIQ SPAQ	Various l Various t Various	levels, sa types, sa s score re	me type me level egions	0.115/ 0.117/ 0.448/	/ 0.081 / / 0.069 / / 0.233 /	0.557 0.416 0.328	0.419 0.325 0.655	9 / 0.402 5 / 0.244 3 / 0.398	2 / 0.9 4 / 0.4 8 / 0.4	906 182 148	0.955 0.905 0.921	5 / 0.925 5 / 0.690 1 / 0.596	/ 0.958 / 0.857 / 0.961
Fable A12 consistent	: Confider and incom	<b>ence sta</b> nsistent	atistics of response	on the fi es.	ne-grai	ned da	taset 1	elease	d by	Zhu	et al.	(2024b	) withi
Datase	ts C	CSIQ (va	rious leve	ls)	CSI	Q (vario	us typ	es)			SPA	Q	
	Cor	nsistent	Incons	istent	Consis	tent	Incons	istent	Co	nsiste	nt	Inconsis	stent
Confide	nce   0.933	3±0.120	0.629±	0.083 0	$0.860 \pm 0$	0.141	).611±	0.098	0.90	0±0.1	125	0.649±0	0.110
MPRNet ( MAXIM ( ht al., 2021	(Zamir e Tu et al., 2 I). For FI	al., 20 et al., 2 2022), M BCNN,	21), HA 021), Ro MPRNet " $q=90$ " 1	AT (Che estorme (Zamir means ti	en et a er (Zan et al., 2 raining	1., 202 nir et 021), I on the	23b), al., 20 DRBN quali	X-Res 022), 1 et (Rua ty facto	torm FBC an et or 90	er (C NN ( al., 2 , and	d im Chen (Jiang 022), "blir	et al., g et al. , and IFA nd" mea	2023a , 2021 AN (Le ns blin
MPRNet ( MAXIM ( et al., 2021 to the qual	(Zamir e Tu et al., 2 I). For FI ity factor	al., 20 et al., 2 2022), M BCNN, :	21), HA 021), Ro MPRNet "q=90" 1	AT (Che estorme (Zamir means tr	en et a er (Zan et al., 2 raining	1., 202 nir et 021), I on the	23b), al., 20 DRBN qualit	X-Res D22), 1 et (Rua ty facto	FBC an et or 90	er (C NN ( al., 2)	d im Chen (Jiang 022), "blir	et al., g et al. , and IFA	2023a , 2021 AN (Le ns blin
MPRNet ( MAXIM ( et al., 2021 to the qual 	(Zamir e Tu et al., ' l). For FF ity factor mage resto uper-resol Denoising PEG comp	al., 20 et al., 2 2022), N BCNN, c pration ta lution pression	21), HA 021), Ro MPRNet "q=90" 1 usk	AT (Che estorme (Zamir means tr emoval	en et a er (Zan et al., 2 raining	I., 202 nir et 021), I on the Image 1 SwinIR SwinIR SwinIR	23b), al., 20 DRBN qualit restora , HAT , MPR , FBC	X-Res 22), 1 et (Rua ty factor tion mo , X-Res Net, Res NN ( $q=$	bdels torma bdels torma estorr 90), 1	er (C NN ( al., 2) , and er ner, X FBCN	(Jiang (Jiang (22)), "blir (-Rest IN (bl	tormer lind)	2023a , 2021 AN (Le ns blin
MPRNet ( MAXIM (' et al., 2021 to the qual to the qual In S D JI	(Zamir e Tu et al., 7 1). For FF ity factor mage restor uper-resol Denoising PEG comp foction deb	al., 20 et al., 2 2022), N BCNN, c pration ta lution pression plurring	21), HA 021), Ro MPRNet "q=90" 1 ask	AT (Che estorme (Zamir means tr emoval	en et a er (Zan et al., 2 raining	I., 202 nir et 021), I on the Image I SwinIR SwinIR SwinIR MAXII	23b), al., 20 DRBN qualit restora , HAT , MPR , FBC M, MP	X-Res 222), 1 et (Rua ty facto tion mo , X-Res .Net, Ro NN (q= RNet, F	bdels torma bdels torma estorn 90), 1 Restor	er (CNN ( al., 2), and er mer, X FBCN rmer	La Ima Chen (Jiang (022), "blir "blir C-Rest	tormer lind)	2023a , 2021 AN (Le ns blin
MPRNet ( MAXIM ( et al., 2021 to the qual ito the qual In S D JI M D JPEG com	(Zamir e Tu et al., ' 1). For FF ity factor mage resto uper-resol penoising PEG comp fotion deb Defocus de	al., 20 et al., 2 2022), N BCNN, c pration ta lution pression plurring blurring artifact	21), HA 021), Ro MPRNet "q=90" 1 ask artifact re removal	xT (Cha estorme (Zamir means tr emoval	er (Zan et al., 2 raining	Image 1 Image 1 Image 1 SwinIR SwinIR SwinIR MAXII DRBN	estora , HAT , FBC M, MPR , FBC M, MPR et, IFA	X-Res 222), 1 et (Rua ty factor tion mo , X-Res NN (q= RNet, F N, Rest	torm FBC an et or 90 odels torma estorr 90), I Restor corme	er (C NN ( al., 20, , and er mer, X FBCN tmer r	a ima Chen (Jiang (Jiang (222), "blir C-Rest IN (bl	tormer lind)	2023æ , 2021 AN (Le ns blin
MPRNet ( MAXIM (' et al., 2021 to the qual	(Zamir e Tu et al., 2 1). For FF ity factor mage restor uper-resol Denoising PEG comp fotion deb Defocus de mpression	al., 20 et al., 2 2022), N BCNN, coration ta lution pression blurring blurring artifact	21), HA 021), Ro MPRNet " $q=90$ " 1 Isk artifact re removal	AT (Cheestorme (Zamir means the emoval	en et a er (Zan et al., 2 raining	In 202 ir et 2021), I on the Image 1 SwinIR SwinIR SwinIR MAXII DRBN The image observed a detail, par appears to details and Comparing It retains 1	23b), al., 2 DRBN qualit restora , HAT , MPR , FBC M, MP et, IFA s depict a mix of A, there es depict a mix of A, there es depict a mix of ticularly be free colors.	X-Res (22), 1 et (Rua ty factor tion mo , X-Res (Net, Res (Net, F N, Rest two tigers orange and is a slight is a slight from such	proceed to the process of the proces	er (C NN ( al., 2 <sup>j</sup> , and er ner, X FBCN mer r tripes, w ression a tripes, w tripes, wall	d Im. Chen (Jian; 022), "'blir CRest IN (bl IN (bl)) IN (bl) IN (	tormer lind) black strip tural backg oresent, whi ing the shar the tigers, og the integ	2023£ , 2021 AN (L¢ ns blin 
MPRNet ( MAXIM (' et al., 2021 to the qual In S DPEG com JPEG com Image Dablur	(Zamir e Tu et al., ' 1). For FF ity factor mage resto uper-resol benoising PEG comp Action deb Defocus de pression	al., 20 at al., 2 2022), N BCNN, : pration ta lution pression oburring blurring artifact	21), HA 021), Ro MPRNet "q=90" 1 ask artifact re removal	AT (Che estorme (Zamir means th emoval	en et a er (Zan et al., 2 raining	Image of the image	s depict a depict colors, the clarity be free colors.	tion mo tion mo tion mo , X-Res Net, Res NN (q= RNet, F N, Rest two tigers orange and is a blockine noticeable from such	dels torm FBCl an et pr 90 dels torma estorr 90), 1 Restor orme so in the distort i of a grades ti	er (C NN ( al., 2 <sup>1</sup> , and ar, 2 <sup>1</sup> , and er ner, X FBCN mer r r tith oran tripes, w ession age e image in the tripes, manual tripes, manual tr	d Im. Chen (Jiang 022), "blir "blir C-Rest IN (bl IN (bl age and with a na c, reduc c, reduc ture of initiality B, Imagg	black strip tural background background black strip tural background the tight background the	2023a , 2021 AN (Le ns blin 
MPRNet ( MAXIM (' et al., 2021 to the qual	(Zamir e Tu et al., ( 1). For FF ity factor mage restor uper-resol penoising PEG comp fotion deb befocus de mpression e A (SwinIR	al., 20 al., 20 al., 2 2022), N BCNN, : oration ta lution pression purring artifact	21), HA 021), RA 021), RA MPRNet "q=90" 1 isk artifact re removal	AT (Che estorme (Zamir means the emoval emoval	in et al., 2 raining	Image 1 Image 1 Image 1 SwinIR SwinIR SwinIR SwinIR MAXII DRBN The image other with In Image detail, par appears to details and Comparing It retains 1 slight com	s depict and a lar pression s colors, there a mix of A, mPR a mix of A, mPR t, IFA s depict a mix of A, there t, and a lar A, there the image he image he poster g clarity a t the two of a high	X-Res (22), 1 et (Rua ty factor tion moo , X-Res NN (q= RNet, F N, Rest two tigers orange and is a slight e blockine noticeable from such ty of Image an outdoo ge poster i is a slight content, p and units.	dels torm FBC: an et pr 90 dels torme estorr 90), 1 Restor comme , one w white s compr 4 (), one w compr 4 (), one w compr 2 (), one w compr 4 (), one w	e with a ckefgroun istortion, market with a ckefgroun istortion and a ckefgroun a ckeff of the c	d Im. Chen (Jiang 022), "blir C-Rest IN (bl C-Rest IN (bl IN (bl) IN (b	et al., g et al., g et al., g et al. ad'' mea tormer lind) black strip- tural backgro- present, while alther the tigers. g the share in the share in the share in the share in the share ditioning ur ing the share on the text e free from t retains the or to Image	2023; , 202 AN (L, ns blin 

Table A10: Relationships between the comparison reasoning task and instant rating task.
Overall, these two tasks are beneficial to each other.

Figure A8: Qualitative results of detailed comparison reasoning on model-processed images.



350	Table A14: Question pool and answer pool of <i>distortion identification</i> ta
351	
352	# Question / Answer
353	Q: What are the primary degradation(s) observed in the evaluated image?
254	A: The primary degradation(s) in the evaluated image is/are {}.
504	2 Q: What distortion(s) are most apparent in the evaluated image?
355	A: The most apparent distortion(s) in the evaluated image is/are {}
356	3 Q: identify the chief degradation(s) in the evaluated image.
257	O: Pinpoint the foremost image quality issue(s) in the evaluated image.
	A: The foremost image quality issue(s) is/are {}.
58	Q: What distortion(s) stand out in the evaluated image?
359	A: The distortion(s) that stand out is/are {}.
360	Q: What distortion(s) are most prominent in the evaluated image?
61	A: The most prominent distortion(s) is/are {}.
301	7 Q: What critical quality degradation(s) are present in the evaluated image?
362	A: The critical quality degradation(s) presented is/are $\{\}$ .
363	$4$ The most significant distortion(s) in the evaluated image is/are $\int$
864	O: What distortion(s) most detrimentally affect the overall quality of the evaluated image?
T	9 A: The distortion(s) that most detrimentally affect the overall quality is/are {}.
65	Q: Determine the most impactful distortion(s) in the evaluated image.
66	A: The most impactful distortion(s) in the evaluated image is/are $\{\}$ .
67	Q: Identify the most notable distortion(s) in the evaluated image's quality.
<u> </u>	A: The most notable distortion(s) in the evaluated image's quality is/are {}.
68	2. Q: What distortion(s) most significantly affect the evaluated image?
69	A: The distortion(s) that most significantly affect the evaluated image is/are {}.
70	13 Q: Determine the leading degradation(s) in the evaluated image.
71	A: The leading degradation(s) is/ate $\{\}$ . O: What distortion(s) are most prominent when examining the evaluated image?
1	A: The most prominent distortion(s) is/are {}
72	O: What distortion(s) are most evident in the evaluated image?
373	<sup>15</sup> A: The most evident distortion(s) in the evaluated image is/are {}.
27/	Q: What quality degradation(s) are most apparent in the evaluated image?
)/+	<sup>10</sup> A: The most apparent quality degradation(s) is/are {}.
75	17 Q: In terms of image quality, what are the most glaring issue(s) with the evaluated image?
76	A: The most glaring issue(s) with the evaluated image is/are {}.
77	Q: What are the foremost distortion(s) affecting the evaluated image's quality?
70	A: The foremost distortion(s) affecting the evaluated image's quality is/are {}.
678	$\frac{Q}{A}$ The most critical distortion(s) in the evaluated image.
79	A: The most efficient distortion(s) is/after $\{\}$ . (): In the evaluated image, what distortion(s) are most detrimental to image quality?
80	$20$ A: In the evaluated image, $\{\}$ is/are the most detrimental distortion(s) to image quality
04	O: What are the most severe degradation(s) observed in the evaluated image?
01	A: The most severe degradation(s) is/are $\{\}$ .
82	22 Q: What are the leading distortion(s) in the evaluated image?
83	A: The leading distortion(s) in the evaluated image is/are $\{\}$ .
0.1	23 Q: What are the most critical image quality issue(s) in the evaluated image?
004	A: The most critical image quality issue(s) in the evaluated image is/are {}.
85	24 Q: What distortion(s) most notably affect the clarity of the evaluated image?
386	A: The distortion(s) that most notably affect the clarity is/are {}.

Table A15: Question pool of assessment reasoning task.

1389	# Question	
1390	1 Could you assess the overall quality of the image and elaborate on your evaluation?	
1301	2 How would you rate the image's quality, and what factors contribute to your assessment?	
1551	3 Can you provide a detailed evaluation of the image's quality?	
1392	4 Please evaluate the image's quality and provide your reasons.	
1393	5 How do you perceive the quality of the image, and what aspects influence your judgment?	
	6 Offer an assessment of the image's quality, highlighting any strengths or weaknesses.	
1394	7 What is your opinion on the quality of the image? Explain your viewpoint.	
1395	8 Assess the quality of the image with detailed reasons.	
1000	9 How does the image's quality impact its overall effectiveness or appeal?	
1396	10 Evaluate the image's quality and justify your evaluation.	
1397	11 How about the overall quality of the image, and why?	
1202	12 Provide a thorough evaluation of the image's quality.	
1390	13 Examine the image's quality by considering factors influencing its clarity.	
1399	14 Analyze the image's quality, and detail your findings.	
1400	15 Provide a comprehensive assessment of the image's quality, including both strengths and areas for improvem	ent.
1-100	16 Assess the image's quality from a professional standpoint.	
1401	17 Evaluate the image's clarity and explain how it contributes to the overall quality.	
1402	18 How would you rate the overall quality of the image, and why?	
1400	19 What is your opinion on the image's quality? Elaborate on your evaluation.	
1403	20 Evaluate the quality of the image and provide a comprehensive explanation.	

1405	
1406	
1407	Table A16: Question pool and answer pool of <i>instant rating</i> task.
1408	# Question / Answer
1409	Q: Which image do you believe has better overall quality: Image A or Image B?
1410	A: I believe Image {} has better overall quality.
411	2 Q: Determine which image exhibits higher quality between image A and image B. A: In my assessment Image {} exhibits higher quality
412	Q: Compare the general quality of Image A and Image B, and state your preference.
413	<sup>5</sup> A: My preference leans towards Image {} to have better general quality.
414	Q: In your opinion, which image demonstrates superior quality: Image A or Image B?
415	A. In my opinion, mage $\{\}$ demonstrates superior quantity. Or which of the two images $\{$ made $\{$ are made $B$ , do you consider to be of better quality?
116	5 4. Loonsider Image A to be of better quality.
410	Q: Evaluate the quality of Image A and Image B, and decide which one is superior.
417	• A: I conclude that Image {} is superior.
418	2 Q: Between Image A and Image B, which image do you think has better quality overall?
419	A: I think Image {} has better quality overall.
120	Q: Determine which image, Image A or Image B, you perceive to have better quality.
420	A: I determine that image {} has belief quality. O: Assess the quality of Image A and Image B and choose the one you believe is superior
421	9 A: I choose Image {} to be superior in terms of quality.
422	Q: Which image stands out to you as having better quality: Image A or Image B?
423	<sup>10</sup> A: Image {} stands out as the superior choice in terms of quality.
101	1 Q: Can you compare the quality of Image A and Image B and decide which one is better?
424	A: I find Image {} to be better after comparing the quality of both.
425	<sup>12</sup> Q: Decide which image, Image A or Image B, you think possesses higher quality.
426	A: I decide that image $\{\}$ possesses inglicit quality.
497	A: Upon evaluation I select Image {} as the one with better quality.
400	. O: Which of the two images, Image A or Image B, appears to have superior quality to you?
428	<sup>14</sup> A: To me, Image {} appears to have superior quality.
429	Q: Compare the quality of Image A and Image B, and determine which one you prefer.
430	A: My preference leans towards Image {} after comparing the quality.
431	Q: Make a judgment on which image, Image A or Image B, you consider to be of better quality.
432	7. Q: Between Image A and Image B, which image do you perceive to have better quality overall?
100	<sup>17</sup> A: I perceive Image {} to have better quality overall.
433	2. Assess the quality of Image A and Image B, and indicate which one you find to be better.
434	A: I find Image {} emerges as the better option with superior quality.
435	19 Q: Which image, Image A or Image B, do you think displays better quality when compared? A: When compared Image {} displays better quality
436	Q: Differentiate between Image A and Image B in terms of overall quality and decide which one is superior.
437	A: Image {} differentiates itself with superior quality.
438	
439	Table A17: Question pool of <i>comparison reasoning</i> task.

ŀ

1440	# Question
1441	1 Compare the overall quality of Image A with Image B and provide a comprehensive explanation.
1442	2 Which image has better visual quality, Image A or Image B? Can you explain the comparison results?
443	3 Evaluate the general visual appeal and quality of both Image A and Image B, and elaborate on which one excels.
	4 Discuss the overall impression and quality of Image A versus Image B, and justify your assessment.
1444	5 Compare the overall quality between Image A and Image B, and justify your comparison results.
1445	6 Assess the overall visual quality of Image A and Image B, discussing which one delivers a more compelling visual quality.
446	7 Which image demonstrates higher overall quality, Image A or Image B? Please provide detailed reasoning for your evaluation.
1/1/7	8 Analyze the overall quality of both image A and image B, and explain which image stands out. 9 Compare the paravised quality of Emage A with Image B, providing incidely into their representive strengths and weaknesses
1447	5 Compare the perceived quarky of image A with image B, providing insigns into their respective strengths and weaknesses. 10 Discuss the visual quality of Image A and Image B, and elaborate on which one anonears more annealing.
448	11 Can you evaluate the overall quality in both Image A and Image B, and explain which one is superior?
1449	12 Compare the overall visual impact and impression of Image A versus Image B, and justify your assessment of their quality.
450	13 Which image exhibits higher overall quality: Image A or Image B? Please explain your reasoning.
151	14 Evaluate the visual quality in Image A and Image B, providing insights into their comparative strengths.
1431	15 Compare the overall quality between Image A and Image B, and discuss which one appears more appealing.
452	16 Assess the visual quality of both Image A and Image B, and explain which one is better.
453	<ol> <li>Which image demonstrates superior quality: Image A or Image B? Please elaborate on your evaluation.</li> <li>Discuss discussed in the superior discussion of the super</li></ol>
151	18 Discuss the overall impression of image A versus image B, and justify your assessment of their comparative quality.
1434	15 Compare the visual quarky or image A with image B, providing detailed insights into their respective strengths and weaknesses. 20 Evaluate the overall unaity of Image A and Image B, and explain which one has higher multity.
1455	20 Enaute die oronan quarty of Enage is and anage of and oppaint which the has inglier quarty.
1456	
1457	



Figure A10: **Qualitative results** on *assessment reasoning* task in the full-reference setting. The two images from top to down are the reference image and evaluated image, respectively.



![](_page_28_Figure_2.jpeg)

![](_page_29_Figure_1.jpeg)

![](_page_29_Figure_2.jpeg)

![](_page_30_Figure_1.jpeg)

Figure A13: **Qualitative results** on *comparison reasoning* task in the non-reference setting. The two images from top to down are Image A and Image B, respectively.

![](_page_31_Figure_1.jpeg)