Q-ALIGN: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels

Haoning Wu [♠♡1] **Zicheng Zhang** [♠2] **Weixia Zhang** [2] **Chaofeng Chen** [1]
Liang Liao [1] **Chunyi Li** [2] **Yixuan Gao** [1 2] **Annan Wang** [1] **Erli Zhang** [1]
Wenxiu Sun [3] **Qiong Yan** [3] **Xiongkuo Min** [2] **Guangtao Zhai** [◇2] **Weisi Lin** [◇1]

Abstract

The explosion of visual content available online underscores the requirement for an accurate machine assessor to robustly evaluate scores across diverse types of visual contents. While recent studies have demonstrated the exceptional potentials of large multi-modality models (LMMs) on a wide range of related fields, in this work, we explore how to teach them for visual rating aligning with human opinions. Observing that human raters only learn and judge **discrete text-defined levels** in subjective studies, we propose to emulate this subjective process and teach LMMs with text-defined rating levels instead of scores. The proposed **Q-ALIGN** achieves state-of-the-art accuracy on *image quality assessment* (IQA), *image aesthetic assessment* (IAA), as well as *video quality assessment* (VQA) under the original LMM structure. With the syllabus, we further unify the three tasks into one model, termed the **ONEALIGN**. Our experiments demonstrate the advantage of discrete levels over direct scores on training, and that LMMs can learn beyond the discrete levels and provide effective finer-grained evaluations. Code and weights will be released.

1. Introduction

There is always a need to score an image. From the early focus on factors related to compression, transmission, and image processing (Sheikh et al., 2005), to directly addressing user-generated content (Tu et al., 2021a) (*e.g.* photos and videos taken with smartphones (Fang et al., 2020)), and moving on to the recently popular AI-generated content (Li et al., 2023), at every stage, accurately evaluating visual content remains an indispensable need to the computer vision field. To address this need, from handcraft approaches (Mittal et al., 2013; 2012) to deep-neural-network-based methods (Talebi & Milanfar, 2018; Zhang et al., 2020; Ke et al., 2021), the endeavor to improve the accuracies of visual evaluators never stops. Nevertheless, while existing methods can already achieve remarkable accuracies on specific datasets by regressing from the mean opinion scores (MOS), the complicated factors that affect the final score in contrast with the limited capacity of these methods have resulted in their poor out-of-distribution (OOD) generalization abilities. This makes them struggle to accurately score novel types of content. Moreover, they usually experience compromised performance while handling different scoring scenarios (*e.g. mixing multiple datasets*) together, making it challenging to train a unified model for different situations.

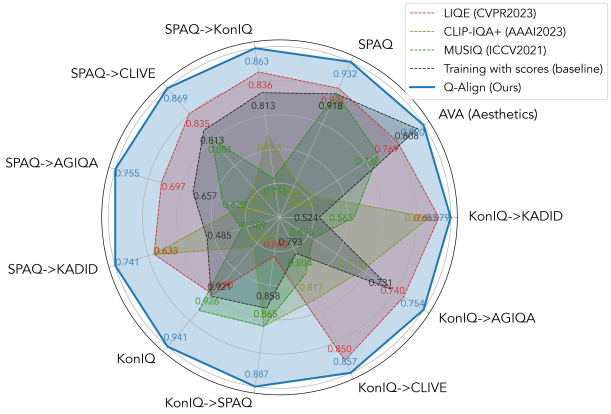


Figure 1. The **Q-ALIGN** (*training LMMs with text-defined levels*) in comparison with its baseline (*training LMMs with scores*) and existing state-of-the-arts, showing exceptional improvements especially on *cross-set* settings. Metrics are (SRCC+PLCC)/2.

In contrast, recently emerging large multi-modality models (LMMs) have shown very strong background knowledge on a wide range of visual and language disciplines. They can well understand high-level visual contents (Liu et al., 2023a; Ye et al., 2023a), and effectively perceive low-level visual attributes (Zhang et al., 2023a), and more importantly possess reasoning ability benefited from their strong language decoder (Liu et al., 2023c). While these abilities are proved fundamental to a more accurate and robust visual

♠ Equal contribution. ♡ Project Lead. [1]Nanyang Technological University [2]Shanghai Jiao Tong University [3]Sensetime Research. Correspondence to: Guangtao Zhai <zhaiguangtao@sjtu.edu.cn>, Weisi Lin <wslin@ntu.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

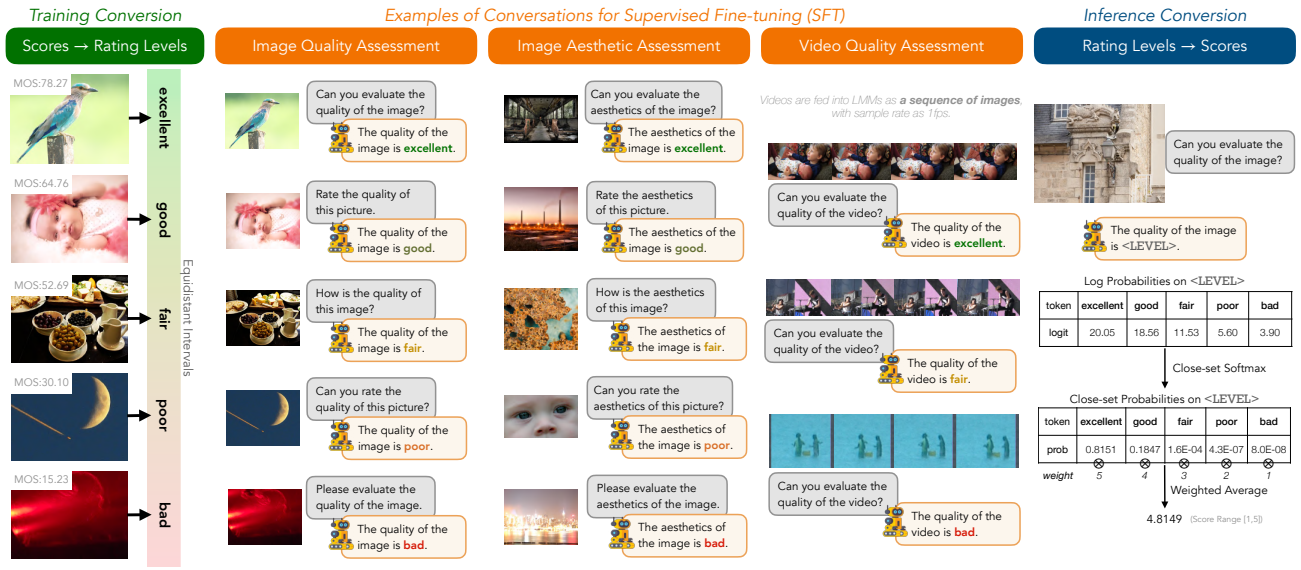


Figure 2. The syllabus of **Q-ALIGN**. Based on the general principle to teach LMMs with text-defined rating levels, we generate the instruction-response pairs by converting **existing score labels** in image quality assessment (IQA), image aesthetic assessment (IAA) and video quality assessment (VQA) datasets. During inference, by simulating the process of collecting mean opinion scores (MOS) from annotators, we extract the close-set probabilities of rating levels and perform weighted average to obtain the LMM-predicted score.

scorer, existing studies (Wu et al., 2023f) have proven that they still fall short on accurately predicting scores that are consistent with human preferences. Therefore, in our study, we investigate the important one last mile for them:

How to teach LMMs to predict scores aligned with human?

To design the most effective syllabus, we reviewed the standard process for collecting MOS from human (itu, 2000): First, organizers need to define several rating levels (*e.g.* '*excellent*', '*fair*', '*bad*') and select examples for each level, aligning human annotators to the standards of each level. Referring to these levels, humans mark their ratings either through a choice button or a grade-guided slider. In other words, **human annotators never learns or marks a specific score** (*e.g. 3.457* in range [1,5]). Instead, these final scores are derived from the distributions of human ratings.

Meanwhile, as observed by recent explorations (Wu et al., 2023f), LMMs have similar *behaviour patterns* to humans while instructed to score: they prefer to **respond with text-defined levels** (*good/poor*); even while explicitly requested to predict numerical scores, the accuracy is significantly lower compared to deriving from levels. Therefore, it might not be optimal to directly tune LMMs to output scores.

Given the above observations, we propose a human-emulating syllabus to teach LMMs for visual scoring (the **Q-ALIGN**), as shown in Fig. 2. **During training**, simulating the process of training human annotators, we convert the MOS values to five text-defined rating levels (itu, 2000) (*excellent/good/fair/poor/bad*), which are further formatted into instruction-response pairs, to conduct visual instruction tuning (Liu et al., 2023b) on LMMs. **During inference**, simulating the strategy to collect MOS from human ratings, we extract the log probabilities on different rating levels, employ softmax pooling to obtain the close-set probabilities of each level. Finally, we get the LMM-predicted score from a weighted average on the close-set probabilities.

While the proposed syllabus requires only existing scores and uses even less information, it has proved far better performance than using scores as learning targets. It reaches **state-of-the-art performance on 12 datasets** of three representative visual scoring tasks with notable improvements: *image quality assessment* (IQA), *image aesthetic assessment* (IAA), and *video quality assessment* (VQA), with especially significant improvements on unseen (OOD) datasets.

Besides achieving state-of-the-art, the proposed **Q-ALIGN** also have two exciting characteristics: **1) Data Efficiency.** It can be competitive with current state-of-the-arts with only 1/5 (IQA) or even 1/10 (IAA) data used. This could be especially useful as data collection is rather expensive for visual scoring tasks. **2) Free Combination of Datasets.** With the strong capacity of LMMs, unlike existing methods that usually face performance drop while mixing datasets (Zhang et al., 2023b), it can freely combine different datasets for training even from different tasks (*i.e. IQA and VQA*), and receive positive performance gain. With this characteristic, we propose the **ONEALIGN**, which combines IQA, IAA and VQA datasets for training. The **ONEALIGN** is exceptionally capable on all three tasks under one unified model, with further enhanced generalization on unseen datasets.

Our core contributions can be summarized as three-fold:

- **An effective syllabus to teach LMMs to score.** Emulating from human experience, the proposed syllabus to train with **discrete levels** is notably more effective than *scores* (**+10%**). Moreover, LMMs can effectively provide finer-grained evaluations under the syllabus.
- **A family of more capable visual assessors.** The proposed **Q-ALIGN** achieves state-of-the-art accuracy and generalization ability on multiple visual assessing tasks. It also proves competitive performance with fewer data used, and can converge with fewer training iterations.
- **A unified model for visual scoring.** With IQA, IAA, and VQA effectively learned independently under the same structure, we further propose **ONEALIGN**, that unifies all three tasks under one model. We hope this may open a new paradigm for visual scoring tasks.

2. Related Works

Image Quality Assessment (IQA). Image quality assessment (IQA) mainly focuses on the impact of distortions and other quality issues in images on human perception. Early IQA algorithms usually operate on handcraft features following the prior knowledge of statistics disciplines (Wang et al., 2004; Mittal et al., 2012; 2013). As distortion diversifies and visual content becomes more complex, data-driven end-to-end deep neural networks are increasingly applied in the IQA field, as represented by NIMA (Talebi & Milanfar, 2018), DBCNN (Zhang et al., 2020), and HyperIQA (Su et al., 2020). Following this path, MUSIQ (Ke et al., 2021) designs a multi-scale input structure that advances the accuracy on IQA via transformers. In recent years, several methods have investigated the vision-language correspondence embedded in CLIP (Radford et al., 2021) to improve generalization ability in IQA. Among them, CLIP-IQA+ (Wang et al., 2022) designs a few-shot learning scheme via CoOp (Zhou et al., 2022), and LIQE (Zhang et al., 2023b) further develops a multitask learning scheme based on CLIP. Nevertheless, they typically rely on visual-text similarity to predict quality scores, which limits their performance to be slightly inferior compared with pure visual methods. Instead, the proposed **Q-ALIGN** can significantly advance state-of-the-arts on IQA, while simultaneously further improving OOD generalization ability.

Image Aesthetic Assessment (IAA). In comparison with IQA, image aesthetic assessment (IAA) (Murray et al., 2012) is a more complicated task for visual scoring. While visual quality is also considered influential to visual aesthetics, the higher-level visual attributes, such as *content, lighting, color, composition* (Kong et al., 2016) are considered more important for IAA. As a result, deep-neural-network-based methods predominate IAA, such as NIMA and MLSP (Hosu et al., 2019). Similar as IQA, VILA (Ke et al., 2023) advances IAA performance by learning vision-language correspondence between images and aesthetic comments (Ghosal et al., 2019) through a joint constrastive and captioning pre-training (Yu et al., 2022). Based on LMMs with rich prior knowledge, the proposed **Q-ALIGN** can remarkably outperform CLIP-based approaches *without* extra pre-training.

Video Quality Assessment (VQA). Named as video *quality* assessment (VQA), the focus of this task is also kind of complicated, that several studies have claimed that scores are not only affected by quality issues, but also contents (Li et al., 2019), and even aesthetics (Wu et al., 2023e). Similar as IQA, while traditional approaches on VQA are typically based on handcraft features, *e.g.* TLVQM (Korhonen, 2019), VIDEVAL (Tu et al., 2021a), and RAPIQUE (Tu et al., 2021b), recent deep-learning-based methods, such as VSFA (Li et al., 2019), BVQA (Li et al., 2022), DisCoVQA (Wu et al., 2023b) and SimpleVQA (Sun et al., 2022), have shown much better performance and more robust OOD generalization. These efforts are further explored by FAST-VQA (Wu et al., 2022; 2023a), which proposes efficient end-to-end training to further advance VQA performance. Nevertheless, while the goal of VQA is similar to IQA (or IAA), the need to input *videos* has hindered methods to tackle this task with the same modeling structure as image scoring approaches. A typical example is the CLIP-based attempts: as CLIP is image-based, though it can achieve good zero-shot VQA capabilities through a frame-by-frame inference (Wu et al., 2023c), training CLIP-based methods on VQA datasets is extremely challenging (Wu et al., 2023d) and performs worse than specially-designed VQA models. In the proposed **Q-ALIGN**, we utilize the language decoder to assemble videos as sequences of frames, so as to unify VQA with IQA/IAA under one structure, outperforming complicated specifically-designed architectures.

LMMs for Visual Scoring. Some recent investigations have discussed the possibilities for adopting Large Multi-modality Models (LMMs) for visual scoring. Namely, the Q-Bench (Wu et al., 2023f) proposes a binary softmax strategy, enabling LMMs to predict quantifiable quality scores by extracting the softmax pooling result on logits of two frequent tokens (*good/poor*). Based on this strategy, the Q-Instruct (Wu et al., 2023g) notices that fine-tuning with text question-answering on related low-level queries can also improve visual scoring abilities of LMMs. Given insights from these studies, we design the **Q-ALIGN** syllabus to systematically emulate the human rating and post-processing in visual scoring. Moreover, we demonstrate that the binary softmax strategy in Q-Bench is a simplified version equivalent to the collection process of MOS values from human ratings. Our experiments prove that with appropriate

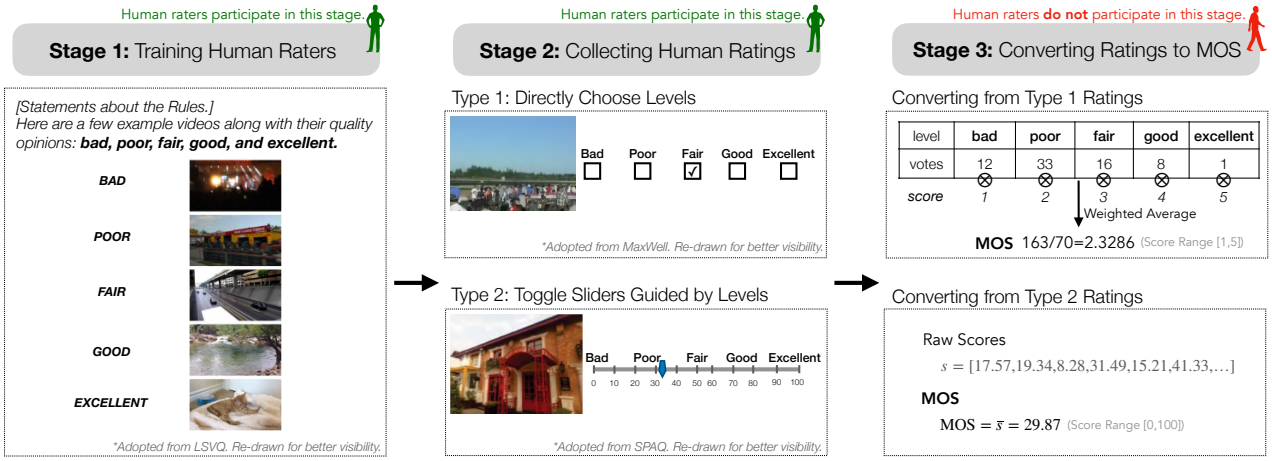


Figure 3. [*Insight 1*] HOW DO HUMANS RATE? Typically, it include three stages: **(1)** Training human raters with text-defined rating levels. Simulating this, we propose the rating-level-based syllabus for LMMs. **(2)** Collecting human ratings. Human raters choose levels (Type 1) or toggle level-guided sliders to score (Type 2), *without directly inputting the score in either way*. **(3)** Converting initial ratings to MOS via weighted average. Following this stage, we propose the probability-based inference for LMMs to predict final scores.

alignment strategies, LMMs can be more capable and robust visual scorers with *the same (and even less) data used*.

3. The Q-ALIGN

In this section, we elaborate on the **Q-ALIGN**. We start with our methodology to teach LMMs with rating levels (Sec. 3.1), and then discuss the proposed conversion strategy between rating levels and scores (Sec. 3.2). Finally, we discuss its unified structure (Sec. 3.3) for images and videos.

3.1. Methodology

3.1.1. [*Insight 1*] HOW DO HUMANS RATE?

To design the syllabus on training LMMs to score, we first review the process of collecting human opinions (Fig. 3). In general, the collection includes three stages as follows:

Stage 1: Training Human Raters. As the standard process for collecting human opinions (itu, 2000), the training process on human raters with the rating rules is vital, including aligning human raters with one or more examples for each **rating level** (Fig. 3 *left*, we take LSVQ (Ying et al., 2021) as an example). During this process, precise quality scores of the examples were *not displayed* to human raters.

Stage 2: Collecting human ratings. After training human raters, the core stage is to collect initial human ratings (Fig. 3 *center*). In general, human raters may provide their opinions in two types: **1)** Directly choose rating levels. **2)** Toggle the slider to generate a score. In either way, human raters do not need to *directly input* the scores to provide their opinions.

Stage 3: Converting human ratings to MOS. As in Fig. 3 *right*, initial ratings are averaged into MOS in visual scoring datasets. Human raters *do not participate* in this stage.

Table 1. [*Insight 2*] HOW DO LMMS RATE? Responses of LMMs on "*Rate the quality of the image*" from 1168 images in LIVE Challenge. LMMs prefer to respond with **qualitative adjectives.**

Model / Frequency	Qualitative Adjectives	Numerical Ratings
Adapter-V2 (Gao et al., 2023)	**96%** (1120/1168)	4% (48/1168)
LLaVA-v1.5 (Liu et al., 2023a)	**100%** (1168/1168)	0% (0/1168)
mPLUG-Owl-2 (Ye et al., 2023b)	**100%** (1168/1168)	0% (0/1168)
InstructBLIP (Dai et al., 2023)	**99%** (1156/1168)	1% (12/1168)
Shikra (Chen et al., 2023)	**100%** (1168/1168)	0% (0/1168)

During all three stages, human raters are **neither trained, nor instructed** to predict a score. This process is adopted because, in everyday life, when asked for an evaluation, people tend to respond with **qualitative adjectives** *(for example, fine, poor, excellent)* rather than **numerical ratings** *(e.g. 8.75, 1.08, 6.54)*. Thus, conducting the visual scoring tasks with rating levels utilizes this *innate ability* of humans (*providing qualitative adjectives*) to minimize their cognitive load, and improve the outcomes of subjective studies.

3.1.2. [*Insight 2*] HOW DO LMMS RATE?

After analyzing the human opinion collection process, we further discover the "*innate ability*" of LMMs. Theoretically, fundamentally designed to understand and generate human-like text, LMMs should share similar behaviour patterns with humans. To validate this, we prompt five LMMs[1] on the instruction as follows, and count their response statistics:

`<img>` *Rate the quality of the image.*

As results shown in Tab. 1, before specific alignment, LMMs predominantly respond with **qualitative adjectives**. Thus, with scores as the learning targets for LMMs, they need to first *formally* learn to output scores, and then learn how to score accurately. To avoid this additional cost, we choose rating levels instead as the targets of **Q-ALIGN**. We study

[1]None of them are explicitly trained for any visual rating tasks.

its advantage than directly training with scores in Tab. 11.

3.2. Conversion between Rating Levels and Scores

Based on the general methodology to teach LMMs with rating levels, we further discuss how to convert the scores in the existing datasets to discrete rating levels *during training*, and how to obtain scores from LMMs *during inference*.

3.2.1. [*Training*] SCORES → RATING LEVELS.

Equidistant Interval Partition. During the training process, we convert the scores into discrete rating levels. Since adjacent levels in human rating are inherently equidistant (either Type 1 or Type 2, see Fig. 3), we also adopt equidistant intervals to convert scores into rating levels. Specifically, we uniformly divide the range between the highest score (M) and lowest score (m) into five distinct intervals, and assign the scores in each interval as respective levels:

$$L(s) = l_i \text{ if } \mathrm{m}+\frac{i-1}{5}\times(\mathrm{M}-\mathrm{m}) < s \leq \mathrm{m}+\frac{i}{5}\times(\mathrm{M}-\mathrm{m}) \tag{1}$$

where $\{l_i|_{i=1}^5\}$ = {*bad, poor, fair, good, excellent*} are the standard text rating levels as defined by ITU (itu, 2000).

Table 2. Precision of training conversion (*Score → Rating Levels*) on the 5 training datasets for **Q-ALIGN**. Metrics are SRCC/PLCC.

Conversion	KonIQ	SPAQ	KADID	AVA	LSVQ
Scores → Levels	0.952/0.961	0.969/0.968	0.979/0.982	0.920/0.930	0.940/0.944

Precision of the Conversion. As the conversion mapping L discussed above is a *multi-to-one* mapping, it unavoidably slightly compromises the ground truth precision. In Tab. 2, we record the conversion precision on 5 datasets used for training **Q-ALIGN**, that all conversion retains around **0.95** linear correlation (PLCC) with the scores. In Appendix Sec. B.2.1, we demonstrate that the **Q-ALIGN** is capable of capturing *finer-grained* differences within each level, even if only the easier *coarse levels* are used for training LMMs.

3.2.2. [*Inference*] RATING LEVELS → SCORES.

After training, we need to convert the rating levels back to scores. Primarily, simulating the post-processing on human ratings (Fig. 3 *right*), we first define the reverse mapping G from text-defined rating levels back to scores, as follows:

$$G : l_i \rightarrow i \tag{2}$$

For instance, *fair* is converted back to score 3, and *bad* to 1.

In human opinion collection (Type 1), the MOS values are calculated via the weighted average of the converted scores and frequencies f_{l_i} for each level: $\mathrm{MOS} = \sum_{i=1}^5 f_{l_i} G(l_i)$. Similarly, for LMMs, we substitute the f_{l_i} with the LMM-predicted probabilities for each level. Given that the predicted <LEVEL> token of LMMs is the probability distribution (denoted as $\mathcal{X}$) on all possible tokens of the language

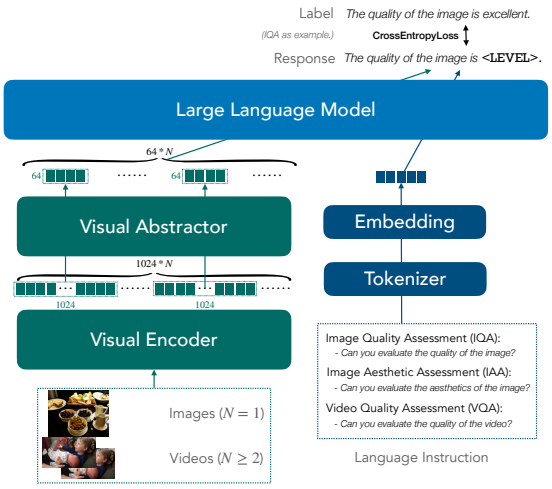


Figure 4. Unified structure of the **Q-ALIGN**. By reducing tokens per image to 64 through the visual abstractor, it unifies images and videos (**as sequences of images**) under one general structure.

model, we conduct a close-set softmax on $\{l_i|_{i=1}^5\}$ to get the probabilities p_{l_i} for each level (p_{l_i} for all l_i sum as 1):

$$p_{l_i} = \frac{e^{\mathcal{X}_{l_i}}}{\sum_{j=1}^5 e^{\mathcal{X}_{l_j}}} \tag{3}$$

and the final predicted scores of LMMs are denoted as

$$\mathrm{S}_{\mathrm{LMM}} = \sum_{i=1}^5 p_{l_i} G(l_i) = i \times \frac{e^{\mathcal{X}_{l_i}}}{\sum_{j=1}^5 e^{\mathcal{X}_{l_j}}} \tag{4}$$

The inference conversion is theoretically equivalent to the MOS collection process from a set of human ratings in levels. Moreover, it represents the general expression form of the binary softmax strategy ($\mathrm{S}_{\text{Q-Bench}} = \frac{e^{\mathcal{X}_{good}}}{e^{\mathcal{X}_{good}}+e^{\mathcal{X}_{poor}}}$) as proposed by Wu et al. (2023f), which can be considered as a simplified version of Eq. 4 with only two rating levels.

3.3. Model Structure

The model structure of the **Q-ALIGN** (Fig. 4) is based on the recently-published open-source LMM, mPLUG-Owl-2 (Ye et al., 2023b), which has proven exceptional visual perception ability as well as good language understanding ability. In the adopted structure, despite the visual encoder to convert images into embeddings, an additional visual abstractor further significantly reduces the token numbers per image (1024 → 64). Under the 4096 context length for LLaMA2 (Touvron et al., 2023), we can feed as much as **61 images** (***3*** *without the abstractor*) together during supervised fine-tuning (SFT). This allows us to input a video **as a sequence of images** to LMM, and unify image (IQA, IAA) and video (VQA) scoring tasks under one structure. The **Q-ALIGN** uses common GPT (Radford et al., 2019) loss, *i.e.* cross-entropy between labels and output logits.

Table 3. **Q-ALIGN** and **FEWSHOT-Q-ALIGN** performance on image quality assessment (IQA). We adopt KonIQ and SPAQ (*both in-the-wild photography*) as training set and evaluate on a wide range of test sets. The cross-set evaluations are labeled with CROSS.

Training Set: **KonIQ$_{train}$**	*→Testing Set:*	**KonIQ$_{test}$**		**SPAQ**^{CROSS}		**LIVE Challenge**^{CROSS}		**AGIQA-3K**^{CROSS}		**KADID-10k**^{CROSS}	
Method	#Training	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIMA (TIP 2018)	7K (70%)	0.859	0.896	0.856	0.838	0.771	0.814	0.654	0.715	0.535	0.532
DBCNN (TCSVT 2020)	7K (70%)	0.875	0.884	0.806	0.812	0.755	0.773	0.641	0.730	0.484	0.497
HyperIQA (CVPR 2020)	7K (70%)	0.906	0.917	0.788	0.791	0.749	0.772	0.640	0.702	0.468	0.506
MUSIQ (ICCV 2021)	7K (70%)	0.929	0.924	0.863	0.868	0.830	0.789	0.630	0.722	0.556	0.575
CLIP-IQA+ (AAAI 2023)	7K (70%)	0.895	0.909	0.864	0.866	0.805	0.832	0.685	0.736	0.654	0.653
LIQE (CVPR 2023)	7K (70%)	0.928	0.912	0.833	0.846	**0.870**	0.830	0.708	0.772	0.662	0.667
FEWSHOT-Q-ALIGN (Ours)	**2K (20%)**	0.903	0.901	0.871	0.860	0.840	0.845	**0.740**	**0.791**	0.607	0.589
Q-ALIGN (Ours)	7K (70%)	**0.940**	**0.941**	**0.887**	**0.886**	0.860	**0.853**	0.735	0.772	**0.684**	**0.674**
Training Set: **SPAQ**	*→Testing Set:*	**KonIQ$_{test}$**^{CROSS}		**SPAQ**		**LIVE Challenge**^{CROSS}		**AGIQA-3K**^{CROSS}		**KADID-10k**^{CROSS}	
Method	#Training	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIMA (TIP 2018)	8.8K (80%)	0.733	0.788	0.907	0.910	0.733	0.785	0.534	0.630	0.399	0.480
DBCNN (TCSVT 2020)	8.8K (80%)	0.731	0.758	0.908	0.913	0.702	0.748	0.459	0.518	0.490	0.508
Fang *et al.* (CVPR 2020)	8.8K (80%)	0.714	0.742	0.908	0.909	0.798	0.762	0.570	0.649	0.381	0.448
MUSIQ (ICCV 2021)	8.8K (80%)	0.753	0.680	0.917	0.921	0.813	0.789	0.564	0.675	0.349	0.429
CLIP-IQA+ (AAAI 2023)	8.8K (80%)	0.753	0.777	0.881	0.883	0.719	0.755	0.577	0.614	0.633	0.638
LIQE (CVPR 2023)	8.8K (80%)	0.826	0.847	0.922	0.919	0.805	0.866	0.672	0.722	0.639	0.627
FEWSHOT-Q-ALIGN (Ours)	**2.2K (20%)**	0.792	0.826	0.909	0.911	0.823	0.834	0.702	0.772	0.685	0.678
Q-ALIGN (Ours)	8.8K (80%)	**0.848**	**0.879**	**0.930**	**0.933**	**0.865**	**0.873**	**0.723**	**0.786**	**0.743**	**0.740**

3.4. Conversation Formats

In this section, we define the conversation formats for each task. Denote the image token as `<img>`, the converted level for the image or video as `<level>`, the exemplar conversation formats for each task are as follows:

Image Quality Assessment (IQA)
#User: `<img>` *Can you evaluate the quality of the image?*
#Assistant: The quality of the image is `<level>`.
Image Aesthetic Assessment (IAA)
#User: `<img>` *How is the aesthetics of the image?*
#Assistant: The aesthetics of the image is `<level>`.
Video Quality Assessment (VQA)
#User: `<img>` *Rate the quality of the video.*
#Assistant: The quality of the video is `<level>`.

The user queries are randomly chosen from a group of *paraphrases* to avoid biases, which shows negligible influence on the final performance. Following Zheng et al. (2023), only the LMM responses (after *#Assistant:*) are supervised.

4. Experiments

4.1. Experimental Settings

In experiments, we set batch sizes as 64 for all IQA/VQA datasets, 128 on IAA datasets, and 256 on **ONEALIGN**. The learning rate is set as $2e-5$, and we train for 2 epochs for all variants, except for FEW-SHOT settings, where we train for 4 epochs to make the models fully converge. All reported performance of **Q-ALIGN** are evaluated on the final weights after training. We conduct training on 4*NVIDIA A100 80G GPUs, and report inference latency on one RTX3090 24G GPU. For images, they are first padded to square and then resized to 448×448. For videos, we sample at rate *1fps*.

4.2. Datasets

IQA datasets. We choose the KonIQ-10k (*in-the-wild*), SPAQ (11K, *in-the-wild*), and KADID-10k (*synthetic*) as training sets to train the **Q-ALIGN** on IQA. Despite evaluating on the test sets on the three training datasets, we also evaluate on four unseen datasets: LIVE Challenge (1.1K, *in-the-wild*), AGIQA-3K (*AI-generated*), LIVE and CSIQ (*both synthetic*) to examine its OOD generalization ability.

IAA datasets. We choose the well-recognized AVA (Murray et al., 2012) dataset to evaluate the aesthetic rating ability of **Q-ALIGN**. Following Hou et al. (2023), we conduct experiments on the OFFICIAL *train-test* split of AVA.

VQA datasets. We choose the largest in-the-wild VQA dataset, LSVQ, with 28K training videos to train the **Q-ALIGN** on VQA. Similar as IQA, we test on two official test sets of LSVQ (LSVQ$_{test}$ and LSVQ$_{1080P}$), and two unseen datasets, KoNViD-1k and MaxWell for OOD evaluation.

4.3. Results on Individual Tasks

4.3.1. IMAGE QUALITY ASSESSMENT (IQA)

For IQA, we first compare the conventional setting where models are trained on a single dataset. As shown in Tab. 3, while CLIP-based methods (CLIP-IQA+ and LIQE) show only comparable or even worse performance on intra-dataset settings than the visual-only state-of-the-art, MUSIQ, the proposed **Q-ALIGN** can notably achieve better accuracy than all visual-only approaches. On cross-dataset settings (OOD generalization), **Q-ALIGN** significantly improves visual-only methods by more than **10%**, and CLIP-IQA+ and LIQE by **8%** and **4%** respectively. In summary, LMM-based **Q-ALIGN** is more competitive under the same data.

Table 4. MIX-DATA experiments for **Q-ALIGN** on image quality assessment (IQA). We label intra-dataset testing sets for each training set combination with gray background, with rest as cross-set settings. Mixing datasets notably improves unseen dataset performance on IQA.

Testing Set:	**KonIQ$_{test}$**		**SPAQ**		**KADID-10k**		**LIVE Challenge**		**AGIQA-3K**		**LIVE**		**CSIQ**	
Training Set:	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
None (mPLUG-Owl2, *Before* **Q-ALIGN**)	0.552	0.489	0.729	0.625	0.572	0.566	0.526	0.538	0.648	0.616	0.521	0.641	0.412	0.393
KonIQ	**0.940**	0.941	0.887	0.886	0.684	0.674	0.860	0.853	0.735	0.772	0.867	0.838	0.700	0.759
SPAQ	0.848	0.879	0.930	0.933	0.743	0.740	0.865	0.873	0.723	0.786	0.861	0.822	0.733	0.781
KonIQ + SPAQ	**0.940**	0.943	**0.931**	**0.933**	0.708	0.692	0.879	0.883	0.727	**0.795**	0.859	0.827	0.767	0.795
KADID	0.668	0.665	0.860	0.854	0.919	0.918	0.702	0.744	0.711	0.712	0.809	0.791	0.756	0.784
KonIQ + SPAQ + KADID	0.938	**0.945**	**0.931**	**0.933**	**0.934**	**0.935**	**0.883**	**0.887**	**0.733**	0.788	**0.870**	**0.840**	**0.845**	**0.876**

Table 5. **Q-ALIGN** performance on video quality assessment (VQA). All methods are trained on the same dataset (LSVQ$_{train}$) and evaluated on two intra-dataset (LSVQ$_{test}$ and LSVQ$_{1080p}$) and two cross-dataset (KoNViD-1k and MaxWell$_{test}$) test sets.

Training Set: **LSVQ$_{train}$**	*→Testing Set:*	**LSVQ$_{test}$**		**LSVQ$_{1080p}$**		**KoNViD-1k^{CROSS}**		**MaxWell$_{test}^{CROSS}$**	
Method	IQA Pre-training?	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
TLVQM (TIP 2019)	✗	0.772	0.774	0.589	0.616	0.732	0.724	–	–
VSFA (ACMMM 2019)	✗	0.801	0.796	0.675	0.704	0.784	0.794	–	–
VIDEVAL (TIP 2021)	✗	0.794	0.783	0.545	0.554	0.751	0.741	–	–
PVQ (CVPR 2021)	✓	0.827	0.828	0.711	0.739	0.791	0.795	0.618	0.634
BVQA (TCSVT 2022)	✓	0.852	0.854	0.772	0.788	0.839	0.830	0.675	0.673
DisCoVQA (TCSVT 2023)	✗	0.859	0.850	0.734	0.772	0.851	0.853	0.704	0.687
SimpleVQA (ACMMM 2022)	✗	0.867	0.861	0.764	0.803	0.840	0.834	0.720	0.715
FAST-VQA (ECCV 2022)	✗	0.876	0.877	0.779	0.814	0.859	0.855	0.720	0.728
Q-ALIGN (Ours) (1fps)	✗	**0.883**	**0.882**	**0.797**	**0.830**	**0.865**	**0.877**	**0.780**	**0.782**
— Ensemble-based Approaches (separately-trained sub-models)									
DOVER (*aesthetic branch* + FAST-VQA, ICCV 2023)	✗	0.886	0.887	0.795	0.830	0.883	0.884	0.748	0.755
Q-ALIGN (Ours) (1fps) + FAST-VQA	✗	**0.899**	**0.899**	**0.818**	**0.850**	**0.895**	**0.897**	**0.779**	**0.784**

Table 6. **Q-ALIGN** performance on image aesthetic assessment (IAA). All methods are trained under the OFFICIAL split setting.

Training Set: **AVA$_{train}$**		*→Testing Set:*	**AVA$_{test}$**	
Method	#Training	Extra Data?	SRCC	PLCC
NIMA (TIP 2018)	236K (92%)	✗	0.612	0.636
MLSP (CVPR 2019)	236K (92%)	✗	0.756	0.757
MUSIQ (ICCV 2021)	236K (92%)	✗	0.726	0.738
MaxViT (ECCV 2022)	236K (92%)	✗	0.708	0.745
CLIP-IQA+ (AAAI 2023)	236K (92%)	✗	0.619	0.586
Aesthetic Predictor (2023)	236K (92%)	✗	0.721	0.723
LIQE (CVPR 2023)	236K (92%)	✗	0.776	0.763
VILA (CVPR 2023)	236K (92%)	✓	0.774	0.774
FEWSHOT-Q-ALIGN (Ours)	**26K (10%)**	✗	0.776	0.775
Q-ALIGN (Ours)	236K (92%)	✗	**0.822**	**0.817**

We further validate that the **Q-ALIGN** can achieve high accuracy with *even less data*. Denoted as **FEWSHOT-Q-ALIGN** in Tab. 3, it reaches comparable performance with existing SOTA IQA approaches by using only 20% images for training, suggesting that the proposed rating-level based approach effectively activates LMM's inherent knowledge.

We further evaluate the mix-dataset scenario for **Q-ALIGN** on IQA in Tab. 4, demonstrating that it can retain and even improve the accuracy on each individual dataset while mixing datasets with different contents (*synthetic* and *in-the-wild*) via simple concatenation, paving the way for the **ONEALIGN** (Sec. 4.4) that unifies different visual scoring tasks. Moreover, each training set combination can improve accuracy on unseen datasets than the pre-alignment baseline.

4.3.2. IMAGE AESTHETIC ASSESSMENT (IAA)

In Tab. 6, we list the results of the **Q-ALIGN** and existing state-of-the-arts on IAA. Compared with IQA, IAA is much more complicated, and the **Q-ALIGN** exhibits far larger advantages with its larger model capacity. It can outperform LIQE by **7%**, Aesthetic Predictor (LAION, 2023) by **10%**. It even significantly improves VILA, which is additionally pre-trained by AVA-Captions, by a notable **6%** margin. Moreover, similar as IQA, the **FEWSHOT-Q-ALIGN** is able to outperform existing IAA methods with only 10% of AVA dataset used for training, further proving the data efficiency of the proposed syllabus on aligning LMMs for scoring.

4.3.3. VIDEO QUALITY ASSESSMENT (VQA)

As listed in Tab. 5, with only sparse frames (*1fps*) as inputs, the **Q-ALIGN** is able to outperform specially-designed VQA approaches with complicated temporal modules and all frames fed into their models. Similar as IQA, it exhibits excellent OOD generalization and surpasses FAST-VQA by 6% on cross-dataset evaluation from LSVQ$_{train}$ to MaxWell$_{test}$ dataset. While **Q-ALIGN** alone can already reach comparable accuracy with DOVER, an approach that ensembles a sparse-frame aesthetic branch with FAST-VQA, its similar ensemble with FAST-VQA proves over 1% advantage to DOVER on all four evaluation datasets. All results suggest that the **Q-ALIGN** can master on VQA with fewer frames as input and no specific design, and still has potential to improve if aligned to rate on more frames in the future.

4.4. The ONEALIGN

Previous evaluations have revealed two exciting abilities of **Q-ALIGN**. First, it reaches state-of-the-art with notable improvements on IQA, IAA, and VQA under one

Table 7. Results of **ONEALIGN** as one unified model for IQA, IAA and VQA, in comparison with single task experts (IQA, IAA, VQA) and partly multi-task experts (IQA+IAA, IQA+VQA, IAA+VQA). LIVE-C abbreviates for LIVE Challenge. Metrics are SRCC/PLCC.

Training / Testing Set	KonIQ	SPAQ	KADID	LIVE-C	AGIQA	LIVE	CSIQ	AVA	$\text{LSVQ}_{\text{test}}$	$\text{LSVQ}_{1080\text{P}}$	KoNViD	MaxWell
Before **Q-ALIGN** (Ye et al., 2023b)	.552/.489	.729/.625	.572/.566	.526/.538	.648/.616	.521/.641	.412/.393	.352/.328	.422/.434	.443/.445	.552/.489	.524/.490
IQA$^{(\overline{KonIQ}+\overline{SPAQ}+\overline{KADID})}$	.938/.945	.931/.933	.934/.935	.883/.887	.733/.788	.870/.840	.845/.876	.208/.228	.755/.757	.680/.718	.799/.806	.682/.694
VQA$^{(LSVQ)}$	.731/.788	.841/.819	.659/.651	.715/.727	.780/.834	.826/.797	.755/.814	.289/.323	.883/.882	.797/.830	.865/.877	.780/.782
IAA$^{(AVA)}$	.574/.603	.662/.653	.536/.547	.685/.636	.750/.792	.770/.740	.527/.596	.822/.817	.624/.600	.515/.511	.717/.681	.659/.648
IQA + VQA	.944/.949	.931/.934	.952/.953	.892/.899	.739/.782	.874/.846	.852/.876	.197/.222	.885/.883	.802/.829	.867/.880	.781/.787
IQA + IAA	.940/.947	.931/.933	.945/.945	.862/.868	.782/.824	.895/.864	.865/.883	.822/.819	.785/.785	.700/.730	.831/.829	.716/.728
IAA + VQA	.640/.664	.740/.732	.626/.632	.703/.669	.769/.819	.794/.769	.558/.628	.822/.819	.886/.885	.800/.834	.874/.884	.776/.781
All (**ONEALIGN**)	.941/.950	.932/.935	.941/.942	.881/.894	.801/.838	.887/.856	.881/.906	.823/.819	.886/.886	.803/.837	.876/.888	.781/.786

Table 8. Epochs to converge for different methods, on KonIQ-10k dataset (IQA). Metrics are (SRCC+PLCC)/2.

Method	*best* (↑)	*Ep1* (↑)	*Ep1 - best* (↑)	#Epochs for *best* (↓)
NIMA (TIP 2018)	0.870	0.650	-0.220	15
CLIP-IQA+ (AAAI 2023)	0.903	0.825	-0.078	12
LIQE (CVPR 2023)	0.920	0.887	-0.033	9
Q-ALIGN (Ours)	**0.942**	**0.931**	**-0.011**	**2**

unified structure. Second, it shows good mix-dataset learning capacity. Moreover, in Tab. 7, we validate that aligning with one task can usually improve on the other tasks (except IQA/VQA→IAA, see Sec. D for more discussions). Given these abilities, we further combine training datasets for the three tasks to train the **ONEALIGN**, the all-in-one visual scorer. As evaluated in Tab. 7, all multi-task variants have shown improved performance than single-task variants. Moreover, the **ONEALIGN** remarkably improves OOD generalization on several unseen datasets: AGIQA$^{+6.8\%\text{SRCC}}$, CSIQ$^{+3.6\%\text{SRCC}}$, LIVE$^{+1.7\%\text{SRCC}}$, KoNViD$^{+1.1\%\text{SRCC}}$. We hope that the **ONEALIGN** can be widely applied to real-world scenarios, pioneering the paradigm shift in this field.

4.5. Cost Analysis

4.5.1. TRAINING COST

As compared in Tab. 8, the **Q-ALIGN** can converge with fewer iterations than existing IQA methods ($bs = 64$ for all), including CLIP-based methods. While existing methods usually need about 10 epochs to reach the best result, the **Q-ALIGN** can outperform all existing methods with only one epoch, and obtain its best results in 2 epochs. With 4*A100 80G GPU, it requires only **9 minutes to converge** on dataset with 10K images, which is highly affordable as it costs less than 2 USD from most cloud GPU providers.

Table 9. Inference latency and throughput of the **Q-ALIGN** on *images* on RTX3090. Larger batch sizes (>64) will cause OOM.

Batch Size	1	2	4	8	16	32	64
Latency (*ms*)	101	154	239	414	757	1441	2790
Throughput (*image/sec*)	9.90	12.99	16.74	19.32	21.14	22.21	22.94

Table 10. Latency and throughput on *videos*. As videos have variable lengths, we set batch size as 1 for them to avoid padding cost.

Video Length (*sec*)	5	7	8	9	10	11	12
Latency (*ms*)	236	315	350	377	430	463	514
Throughput (*video/sec*)	4.24	3.17	2.86	2.65	2.33	2.16	1.95

4.5.2. INFERENCE LATENCY

In Tab. 9 and Tab. 10, we discuss the inference latency of **Q-ALIGN** on images and videos. In one second, it can predict scores on up to **23** images, **4.2** 5s-duration videos, or **1.9** 12s-duration videos on a single RTX3090 GPU. Moreover, we also validate in Appendix Tab. 13 that 4-BIT inference on **Q-ALIGN** has almost identical accuracy. It costs only **5.4GB** vRAM and allows broader application of the scorer.

Table 11. **Q-ALIGN** compared with the variant that use scores (*in* `.2f` *format*) as *training* objective. Metrics are (SRCC+PLCC)/2.

Training Set:	KonIQ				
Testing Set:	KonIQ	SPAQ^{CROSS}	LIVE-C^{CROSS}	AGIQA^{CROSS}	KADID^{CROSS}
Existing SOTA	0.926	0.865	0.850	0.740	0.665
- *Training with Scores*	0.921	0.858	0.793	0.731	0.524
Q-ALIGN (Ours)	**0.941**	**0.887**	**0.857**	**0.754**	**0.679**
Improvement	**+2.2%**	**+3.4%**	**+8.1%**	**+3.1%**	**+29.6%**
Training Set:	SPAQ				
Testing Set:	SPAQ	KonIQ^{CROSS}	LIVE-C^{CROSS}	AGIQA^{CROSS}	KADID^{CROSS}
Existing SOTA	0.921	0.836	0.835	0.697	0.633
- *Training with Scores*	0.918	0.813	0.813	0.657	0.485
Q-ALIGN (Ours)	**0.932**	**0.863**	**0.869**	**0.755**	**0.741**
Improvement	**+1.5%**	**+6.2%**	**+6.9%**	**14.9%**	**+52.8%**

4.6. Ablation Studies

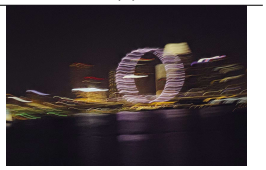
Q-ALIGN *vs* training with scores. In Tab. 11, we compare the **Q-ALIGN** with the variant that directly instructs the LMM to output scores during training. Using the proposed level-based syllabus can lead to in-average **10%** improvements on cross-dataset (OOD) evaluations (*especially* **40%**↑ from SPAQ/KonIQ (*in-the-wild*) to KADID (*synthetic*)) than the score-based syllabus, suggesting that instructing LMMs with their original output styles better inherits their innate visual evaluation potentials. Conversely, the accuracies of score-based alignment cannot even surpass existing state-of-the-art on any settings, unable to effectively inherit the powerful capabilities from the pre-trained LMMs.

Furthermore, we validate that inferring with probabilities helps finer-grained distinction (Fig. 5) and improves final accuracy by **6%** (Tab. 14). More details in Appendix Sec. B.2.

4.7. Qualitative Analysis

In Tab. 12, we visualize the IQA and IAA prediction results of the **ONEALIGN** on two real-world images. Despite the basic ability to judge that (A) > (B) in both *quality* and *aesthetics*, we notice that it can further capture subtle dif-

Table 12. **ONEALIGN** predictions on real-world images, from logits to probabilities, and finally to scores. More in Appendix Sec. C.

	(A)					(B)				
										
<img>										
l_i	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*
$\mathcal{X}^{l_i}_{\text{IQA}}$	18.03	**18.38**	14.63	11.60	9.477	8.953	11.37	15.31	**18.06**	16.59
$p^{l_i}_{\text{IQA}}$	0.409	**0.577**	0.014	0.000	0.000	0.000	0.001	0.050	**0.772**	0.178
$S_{\text{LMM, IQA}}$	**4.3926** (Range: [1,5])					1.8740 (Range: [1,5])				
$\mathcal{X}^{l_i}_{\text{IAA}}$	16.63	**18.17**	15.77	12.13	10.77	9.594	13.13	16.95	**17.67**	14.91
$p^{l_i}_{\text{IAA}}$	0.163	**0.766**	0.069	0.002	0.000	0.000	0.007	0.312	**0.641**	0.040
$S_{\text{LMM, IAA}}$	**4.0879** (Range: [1,5])					2.2861 (Range: [1,5])				

ferences. Though trained with only discrete levels, its *2nd highest level* (underlined) can provide finer-grained evaluations, that the *aesthetics* of (B) is between *fair* and *poor*, while its *quality* lies between *poor* and *bad*. Moreover, the **ONEALIGN** never predicts *1st* and *2nd* highest logits on non-adjacent levels (*e.g. good&poor*), suggesting that the model can inherently understand the ordinals on the levels.

5. Conclusion

In conclusion, our paper marks a significant stride in the realm of visual scoring by instructing Large Multi-modality Models (LMMs) with discrete text-defined levels (*e.g., good, poor*) rather than direct scores (*e.g., 3.45, 1.77*). This syllabus, named the **Q-ALIGN**, achieves remarkable improvements over state-of-the-art IQA, IAA and VQA approaches under one general structure, with exceptional data efficiency. It further unifies all the three tasks under one single model, the **ONEALIGN**. The **Q-ALIGN** unlocks the potential of LMMs in predicting accurate and robust visual scores, pioneering a promising direction for future explorations.

Impact Statement

Aligning to human opinion bias is a general potential impact for all learning-based rating systems. In cross-evaluation settings, **Q-ALIGN** is more generalizable than existing systems, implying that **Q-ALIGN** potentially be less impacted by training data bias. Still, it cannot eliminate this impact, and we will keep focusing on it in the future.

References

Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2000.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., and Qiao, Y. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Ghosal, K., Rana, A., and Smolic, A. Aesthetic image captioning from weakly-labelled photographs. 2019.

Hosu, V., Goldlücke, B., and Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*, 2019.

Hou, J., Lin, W., Fang, Y., Wu, H., Chen, C., Liao, L., and Liu, W. Towards transparent deep image aesthetics assessment with tag-based content descriptors. *IEEE TIP*, 2023.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., and Yang, F. Vila: Learning image aesthetics from user comments with vision-language pretraining, 2023.

Kong, S., Shen, X., Lin, Z., Mech, R., and Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

Korhonen, J. Two-level approach for no-reference consumer video quality assessment. *IEEE TIP*, 2019.

LAION. Aesthetic predictor. https://github.com/LAION-AI/aesthetic-predictor, 2023.

Li, B., Zhang, W., Tian, M., Zhai, G., and Wang, X. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE TCSVT*, 2022.

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

Li, D., Jiang, T., and Jiang, M. Quality assessment of in-the-wild videos. In *ACM MM*, pp. 2351–2359, 2019.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023b.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. Mmbench: Is your multi-modal model an all-around player?, 2023c.

Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12), 2012.

Mittal, A., Soundararajan, R., and Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

Murray, N., Marchesotti, L., and Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pp. 2408–2415, 2012.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Sheikh, H. R., Wang, Z., Cormack, L., and Bovik, A. C. Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality, 2005.

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, June 2020.

Sun, W., Min, X., Lu, W., and Zhai, G. A deep learning based no-reference quality assessment model for ugc videos. In *ACMMM*, pp. 856–865, 2022.

Talebi, H. and Milanfar, P. Nima: Neural image assessment. *IEEE TIP*, 2018.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE TIP*, 30:4449–4464, 2021a.

Tu, Z., Yu, X., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021b.

Wang, J., Chan, K. C. K., and Loy, C. C. Exploring clip for assessing the look and feel of images, 2022.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., and Lin, W. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *ECCV*, 2022.

Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., Gu, J., and Lin, W. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 2023a.

Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., and Lin, W. Discovqa: Temporal distortion-content transformers for video quality assessment. *TCSVT*, 2023b.

Wu, H., Liao, L., Chen, C., Hou, J. H., Zhang, E., Wang, A., Sun, W., Yan, Q., and Lin, W. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *ICME*, 2023c.

Wu, H., Liao, L., Wang, A., Chen, C., Hou, J. H., Zhang, E., Sun, W. S., Yan, Q., and Lin, W. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023d.

Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., and Lin, W. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023e.

Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., and Lin, W. Q-bench: A benchmark for general-purpose foundation models on low-level vision. 2023f.

Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*, 2023g.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., and Huang, F. mplug-owl: Modularization empowers large language models with multimodality, 2023a.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.

Ying, Z., Mandal, M., Ghadiyaram, D., and Bovik, A. Patch-vq: 'patching up' the video quality problem. In *CVPR*, 2021.

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

Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Zhang, W., Yan, H., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., and Wang, J. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023a.

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2020.

Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Table 13. 4-BIT and 8-BIT inference, in comparison with original `fp16` **ONEALIGN**. ^{CR}: unseen datasets. Metrics are SRCC/PLCC. The vRAM cost is for single image. For a 8-sec video, the cost is 5816MB (4-BIT) and 10042MB (8-BIT) respectively.

Testing Set:	vRAM↓	KonIQ	SPAQ	KADID	LIVE-C^{CR}	AGIQA^{CR}	LIVE^{CR}	CSIQ^{CR}	AVA	LSVQ$_{\text{test}}$	LSVQ$_{\text{1080P}}$	KoNViD^{CR}	MaxWell^{CR}
4-BIT (`bitsandbytes`)	5396MB	.937/.947	.931/.933	.939/.939	**.886/.897**	.801/.838	.886/.853	.877/.904	.821/.817	.884/.884	.798/.834	**.877/.888**	.778/.785
8-BIT (`bitsandbytes`)	8944MB	.936/.947	.931/.934	.941/.941	.882/.894	.801/.836	.886/.855	.877/.902	.821/.816	.885/.885	.801/.835	.875/.886	.777/.786
w/o Quantization (`fp16`)	16204MB	**.941/.950**	**.931/.934**	**.941/.942**	.881/.894	**.801/.838**	**.887/.856**	**.881/.906**	**.823/.819**	**.886/.886**	**.803/.837**	.876/.888	**.781/.786**

A. Additional Modeling Details

A.1. Conversation Formats

In this section, we discuss the details on the conversation formats for each task. Denote the image token as `<img>`, the converted level for the image or video as `<level>`, the exemplar conversation formats for each task are as follows:

Image Quality Assessment (IQA)
#User: `<img>` *Can you evaluate the quality of the image?*
#Assistant: The quality of the image is `<level>`.
Image Aesthetic Assessment (IAA)
#User: `<img>` *Can you evaluate the aesthetics of the image?*
#Assistant: The aesthetics of the image is `<level>`.
Video Quality Assessment (VQA)
#User: `<img><img><img><img><img><img><img><img>` *Can you evaluate the quality of the video?*
#Assistant: The quality of the video is `<level>`.

The user queries are randomly chosen from a group of *paraphrases* (e.g. *Rate the quality of the image.*, *How would you rate the aesthetics of the image?*, *How is the quality of the video?*) to avoid biases, which shows negligible influence on the final performance. Following Zheng et al. (2023), only the LMM responses (after *#Assistant:*) are supervised.

A.2. Formulation on Model Structure

Following mPLUG-Owl-2, the model includes a CLIP-ViT-Large visual encoder $\mathbf{E_v}$ with 304M parameters, a visual abstractor $\hat{\mathbf{E}}_\mathbf{v}$ with 82M parameters, and the LLaMA2-7B LLM $\mathbf{D}$ on top of the visual modules with 7.8B (*with the additional multi-way modules from mPLUG-Owl2*) parameters. The input image is first padded to square, and then resized to 448×448. Denote the text embedding layer as $\mathbf{E_t}$, input image as `<img>`, text prompt as t, the model can be formulated as:

$$\mathbf{h_v} = \hat{\mathbf{E}}_\mathbf{v}(\mathbf{E_v}(\texttt{<img>})) \tag{5}$$

$$\mathbf{h_t} = \mathbf{E_t}(t) \tag{6}$$

$$\mathbf{h} = \text{concatenate}(\mathbf{h_v}, \mathbf{h_t}) \tag{7}$$

$$\mathbf{o} = \mathbf{D}(\mathbf{h}) \tag{8}$$

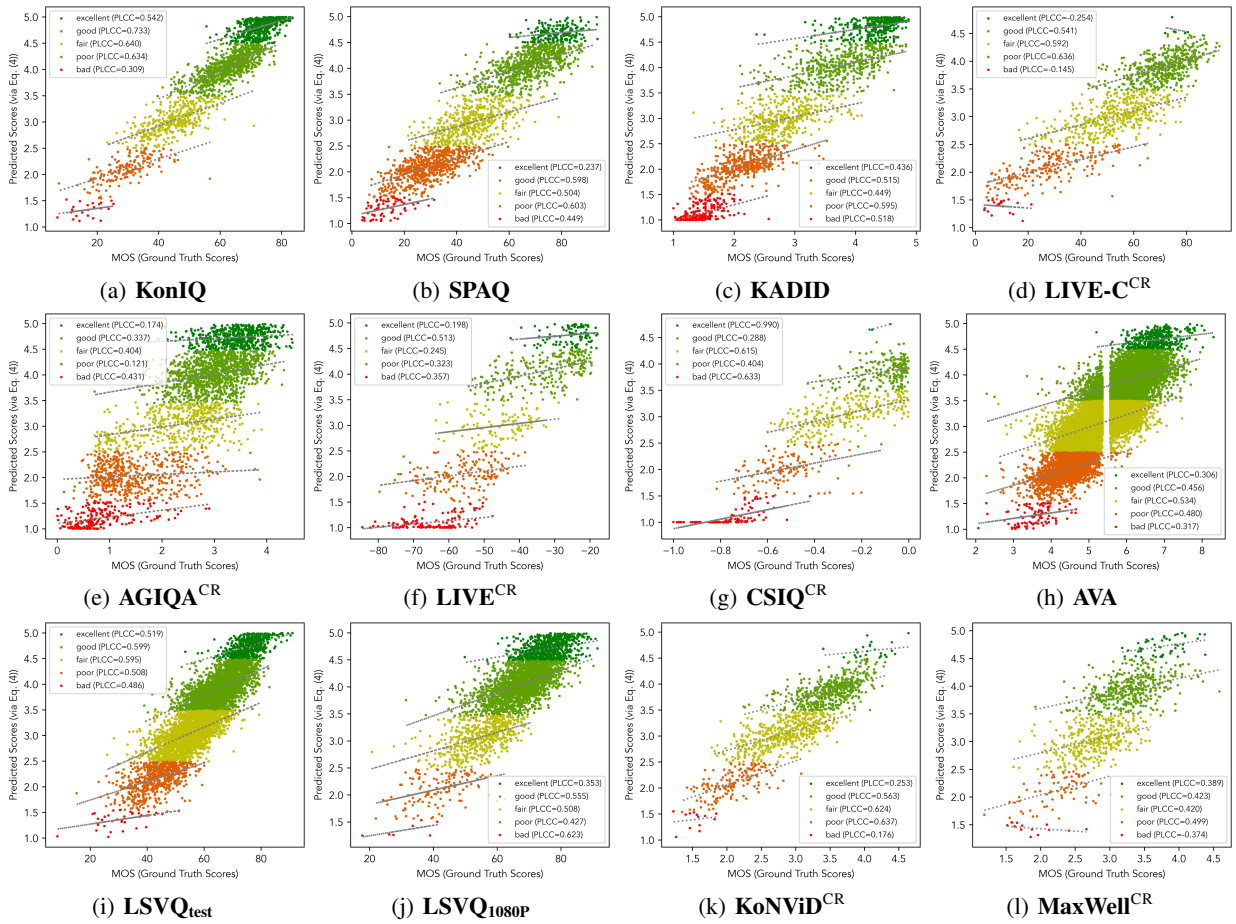
As we do not need the `generation` method of the LMM, the formulation above, where the output logits $\mathbf{o}$ are the final outputs, can define both the training and inference processes for the **Q-ALIGN**. Specifically, during inference, only the input texts before the `<level>` word are fed into the LMM, and henceforth the last position of $\mathbf{o}$ is the desired probability distributions[2], *i.e.* $\mathcal{X} = \mathbf{o}_{N-1}$ (*where N is total output length*). The final output score is obtained as in Eq. 4.

B. Extended Experimental Analysis

B.1. Effects of Quantization (4-BIT&8-BIT)

In Tab. 13, we discuss the impacts of using quantization during inference on the **ONEALIGN**. We notice that even 4-BIT inference only leads to overall 0.2% performance degradation (*and even improves by 0.4% on LIVE-C*), but reduces the vRAM consumption to infer with *bs=1* from 16.2GB to 5.4GB. The reduced memory cost without almost identical performance has greatly broadened the application scenarios of the **ONEALIGN**, that the more powerful and robust LMM-base scorer can be deployed locally on laptops with RTX3060 GPUs, or even a MacBook.

[2]In causal language models (Radford et al., 2019), the $\mathbf{o}_i$ is the prediction for the $(i+1)$-th token.



(a) **KonIQ** (b) **SPAQ** (c) **KADID** (d) **LIVE-C^{CR}**

(e) **AGIQA^{CR}** (f) **LIVE^{CR}** (g) **CSIQ^{CR}** (h) **AVA**

(i) **LSVQ$_{test}$** (j) **LSVQ$_{1080P}$** (k) **KoNViD^{CR}** (l) **MaxWell^{CR}**

Figure 5. Finer-grained distinction abilties of **ONEALIGN**. By grouping its predictions *w.r.t.* max-probability token on `<level>` in 12 evaluation sets, we demonstrate that though trained with as discrete classification, the final predicted score (via Eq. 4) of the **ONEALIGN** can obviously positively correlate with human opinions among finer-grained images/videos even they are *"classified" as the same level*.

Table 14. **ONEALIGN** compared with the variant that use direct levels as *inference* strategy. ^{CR}: unseen datasets. Metrics are SRCC/PLCC.

Testing Set:	KonIQ	SPAQ	KADID	LIVE-C^{CR}	AGIQA^{CR}	LIVE^{CR}	CSIQ^{CR}	AVA	LSVQ$_{test}$	LSVQ$_{1080P}$	KoNViD^{CR}	MaxWell^{CR}
- Inference with Levels	.881/.903	.897/.896	.920/.921	.828/.841	.777/.812	.861/.834	.841/.871	.748/.751	.818/.824	.717/.761	.808/.821	.725/.733
Inference w/ Eq. 4 (Ours)	**.941/.950**	**.931/.934**	**.941/.942**	**.881/.894**	**.801/.838**	**.887/.856**	**.881/.906**	**.823/.819**	**.886/.886**	**.803/.837**	**.876/.888**	**.781/.786**
Improvement	**+6.0%**	**+4.0%**	**+2.3%**	**+6.4%**	**+3.1%**	**+2.8%**	**+4.4%**	**+9.5%**	**+7.9%**	**+11.0%**	**+8.3%**	**+7.5%**

B.2. Effects of the Inference Strategy

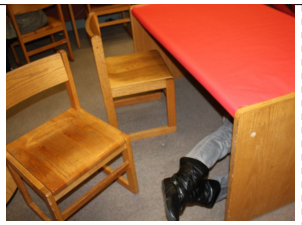
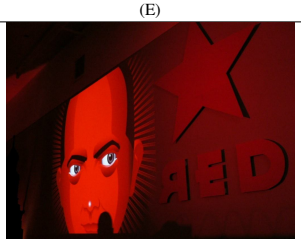
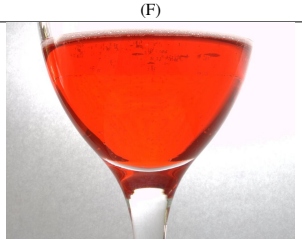
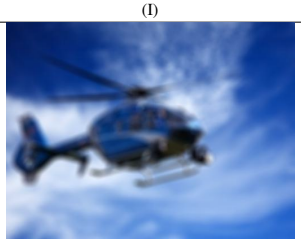
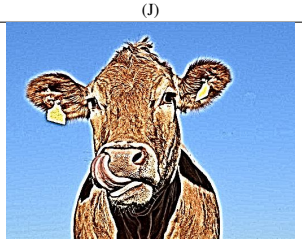
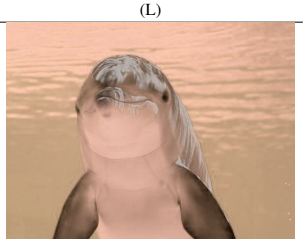
B.2.1. FINER-GRAINED DISTINCTION

In Fig. 5, we discuss the finer-grained distinction ability of the proposed **ONEALIGN**. Specifically, we group the test set samples into five groups, where each group includes samples with highest-probability `<level>` token as the level, *i.e.* *"classified"* as the respective level. While we do not explicitly train the LMM to distinguish beyond the five levels in the **Q-ALIGN** syllabus, the predicted scores shows strong finer-grained alignment with human opinions (*within the same group*).

B.2.2. CONTRIBUTION OF FINER-GRAINED DISTINCTION ON OVERALL ACCURACY

In Tab. 14, we further compare the overall accuracy between **ONEALIGN** and the variant that directly takes the $G(l_i)$ from the highest-probability level as the output score during inference (*i.e.* ignoring its finer-grained prediction). We prove that inferring with probabilities (as in Eq. 4) can significantly improve accuracy on every test set (in average **6%**). Together with evidences in Fig. 5, it suggests that the **Q-ALIGN** syllabus can activate LMM's innate ability to catch finer-grained difference on visual scoring with only discrete rating levels as supervision.

Table 15. Extended **ONEALIGN** predictions on image quality assessment (IQA), from logits to probabilities, and finally to scores. The first eight images ((A) - (H)) are *in-the-wild* images, while others ((I) - (P)) are images with *synthetic distortions*. Zoom in to view details.

	(A)					(B)					(C)					(D)				
																				
<img>																				
l_i	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*
$x_{\text{IQA}}^{l_i}$	7.043	10.289	15.852	**18.688**	16.766	6.7500	13.117	18.969	**19.062**	13.984	9.211	16.188	**19.188**	17.047	11.586	10.086	16.125	**18.859**	16.734	11.891
$p_{\text{IQA}}^{l_i}$	0.000	0.000	0.049	**0.830**	0.121	0.000	0.001	0.474	**0.521**	0.003	0.000	0.043	**0.856**	0.101	0.000	0.000	0.055	**0.844**	0.101	0.001
$S_{\text{LMM, IQA}}$	**1.9277**					**2.4746**					**2.9414**					**2.9531**				
	(E)					(F)					(G)					(H)				
																				
<img>																				
l_i	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*
$x_{\text{IQA}}^{l_i}$	12.500	17.594	**18.406**	15.227	11.008	14.781	**19.359**	17.609	13.172	9.117	15.547	**19.969**	17.297	12.477	8.648	**20.562**	18.093	10.992	9.5088	8.055
$p_{\text{IQA}}^{l_i}$	0.002	0.298	**0.672**	0.028	0.000	0.009	**0.843**	0.146	0.002	0.000	0.011	**0.924**	0.064	0.001	0.000	**0.922**	0.078	0.000	0.000	0.000
$S_{\text{LMM, IQA}}$	**3.2734**					**3.8594**					**3.9453**					**4.9219**				
	(I)					(J)					(K)					(L)				
																				
<img>																				
l_i	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*
$x_{\text{IQA}}^{l_i}$	9.820	10.055	9.070	16.016	**25.062**	7.047	8.375	10.867	16.531	**19.656**	5.559	8.914	15.930	**20.078**	16.516	9.805	13.289	16.781	**17.750**	14.172
$p_{\text{IQA}}^{l_i}$	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	0.042	**0.958**	0.000	0.000	0.015	**0.958**	0.027	0.000	0.008	**0.267**	0.705	0.020
$S_{\text{LMM, IQA}}$	**1.0000**					**1.0420**					**1.9873**					**2.2656**				
	(M)					(N)					(O)					(P)				
<img>																				
l_i	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*	*excellent*	*good*	*fair*	*poor*	*bad*
$x_{\text{IQA}}^{l_i}$	10.883	17.516	**19.062**	15.156	9.203	11.469	17.031	**18.375**	15.391	10.586	15.758	**18.609**	16.641	12.977	9.188	**18.547**	17.625	13.766	11.258	8.781
$p_{\text{IQA}}^{l_i}$	0.000	0.173	**0.811**	0.016	0.000	0.001	0.199	**0.762**	0.039	0.000	0.048	**0.833**	0.116	0.003	0.000	**0.711**	0.283	0.006	0.000	0.000
$S_{\text{LMM, IQA}}$	**3.1582**					**3.1621**					**3.9258**					**4.7031**				

C. Extended Qualitative Analysis

In Tab. 15, we visualize more qualitative results on IQA (*logits, probabilities and scores*) from **ONEALIGN**, on both *in-the-wild* images ((A)-(H)) and images *with synthetic distortions* ((I)-(P)). Under one single model (*instead of separately trained for in-the-wild and synthetic images*), it can distinguish both common in-the-wild degradation, *e.g.* under-exposure (A,B), low sharpness/resolution (A,D), realistic noise (C), as well as the synthetic distortions such as artificial blur (I), artifacts (J), gaussian noises (K). It can also distinguish distortion levels (*e.g.* strong gaussian noise on (K) *vs* weak gaussian noise on (M)). We hope that the **ONEALIGN** can work as a robust quality scorer in real-world scenarios.

D. Relations among IQA, IAA, and VQA

In Tab. 7, we notice that **single-task** fine-tuned model variants present different impacts on other tasks, discussed as follows:

1. Both image-based (IQA, IAA) variants can notably improve accuracy on VQA, suggesting that the image-related issues (*clarity, brightness, noises*) are highly influential in VQA. This conclusion aligns with the observed non-negligible mix-task gain of the **ONEALIGN** on all four VQA evaluation sets: $\text{LSVQ}_{\text{test}}^{+0.4\%\text{PLCC}}$, $\text{LSVQ}_{\text{test}}^{+0.8\%\text{PLCC}}$, $\text{KoNViD}^{+1.1\%\text{PLCC}}$, $\text{MaxWell}^{+0.5\%\text{PLCC}}$, in comparison with **Q-ALIGN**-VQA.

2. While the aesthetic (IAA) variant can generally improve accuracy on quality evaluation (IQA/VQA), the IQA/VQA *in turn degrades* the aesthetic evaluation ability of the LMM. This relation might suggest that quality evaluation considers only subset of issues (*clarity, color, brightness*) that are considered in aesthetic evaluation, while there are still many aesthetic-related factors not considered in IQA (*composition, theme, etc*). While it does not affect the rationality of multi-tasking between quality and aesthetic evaluation, we believe this is an interesting finding to point out.

E. Further Clarifications

E.1. Novelty of Q-ALIGN

While the **Q-ALIGN** architecture is based on existing models, we would like to emphasize our technical contributions as follows:

1. It broadens the scope of LMMs. While alignment studies for LMMs mainly focuses on situations with text outputs, this work presents the first attempt for them to quantitatively score/evaluate. While no extra plugin structure required, the alignment could be seamlessly merged into general LMM fine-tuning.

2. It presents an alternative methodology for scoring tasks. Q-Align is a rare method to use classification instead of regression as training objective for scoring tasks, and proves its effectiveness against existing regression-based approaches.

E.2. What does the Q of Q-ALIGN refer to?

The **Q** here mainly refers to **Quantitative Evaluation**, as it is a study to align LMMs to provide quantitative evaluation (scores, instead of text outputs). It also implies the broad quality assessment task (a long-existing domain for visual scoring, and the main task **Q-ALIGN** tackles with).

E.3. Social Impact

Aligning to human opinion bias is a general potential impact for all learning-based rating systems. In cross-evaluation settings, **Q-ALIGN** is more generalizable than existing systems, implying that **Q-ALIGN** potentially be less impacted by training data bias. Still, it cannot eliminate this impact, and we will keep focusing on it in the future.

F. Open-Source Commitments

The **Q-ALIGN** does not contain any additional human subjective studies. Henceforth, to promote the open-source community to expand this syllabus to more related tasks, we will open-source all training data (converted from existing datasets), training and inference code, and pre-trained weights under MIT License, upon the acceptance of our manuscript.