Adaptive Prompt Learning for Blind Image Quality Assessment with Multi-modal Mixed-datasets Training

Yan Zhong
School of Mathematical Sciences, State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University.
Beijing, China
zhongyan@stu.pku.edu.cn

Xinping Zhao
School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
Shenzhen, China
24B351003@stu.hit.edu.cn

Li Zhang
Hefei Institute of Physical Science, Chinese Academy of Sciences, University of Science and Technology of China
Hefei, China
zanly@mail.ustc.edu.cn

Xinyuan Song
Department of Computer Science and Engineering, University of Texas at Arlington
Arlington, Texas, the United States
xxs1853@mavs.uta.edu

Tingting Jiang*
National Engineering Research Center of Visual Technology, State Key Laboratory of Multimedia Information Processing, School of Computer Science, National Biomedical Imaging Center, Peking University
Beijing, China
ttjiang@pku.edu.cn

Abstract

Due to the high cost and small scale of Image Quality Assessment (IQA) datasets, achieving robust generalization remains challenging for prevalent Blind IQA (BIQA) methods. Traditional deep learning-based methods emphasize visual information to capture quality features, while recent developments in Vision-Language Models (VLMs) demonstrate strong potential in learning generalizable representations through textual information. However, applying VLMs to BIQA poses three major **Challenges**: **(1)** How to make full use of the multi-modal information. **(2)** The prompt engineering for appropriate quality description is extremely time-consuming. **(3)** How to use mixed data for joint training to enhance the generalization of VLM-based BIQA model. To this end, we propose a Multi-modal BIQA method with prompt learning, named MMP-IQA. For **(1)**, we propose a conditional fusion module to better utilize the cross-modality information. By jointly adjusting visual and textual features, our model can capture quality information with a stronger representation ability. For **(2)**, we model the quality prompt's context words with learnable vectors during the training process, which can be adaptively updated for superior performances. For **(3)**, we jointly train a linearity-induced quality evaluator, a relative quality evaluator, and a dataset-specific absolute quality evaluator. In addition, we propose a dual automatic weight adjustment strategy to adaptively balance the loss weights between different datasets and among various losses within the same dataset. Extensive experiments illustrate the superior effectiveness and generalization ability of MMP-IQA.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → *Image processing*; *Visual inspection*; *Image representations*; *Supervised learning by regression*.

Keywords

Image Quality Assessment, Generalization, Multi-modal Fusion, Prompt Learning, Mixed Data Training

1 Introduction

Image quality assessment (IQA) plays a crucial role in optimizing visual experiences across different domains, including image denoising [60], restoration [10], and generation [12], which aims to develop algorithms that can accurately predict image quality scores consistent with subjective human ratings (i.e., Mean Opinion Scores, MOS). Based on the availability of reference information, existing IQA methods can be classified into Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and Blind IQA (BIQA) [75]. In this paper, we focus on the BIQA methods, as their reference-free nature allows a wider applicability [15, 54, 55, 86].

Inspired by the success of Deep Learning (DL) in fusing discriminative features in visual domains [21, 37, 55], numerous studies have sprung up on DL-based BIQA methods for authentically distorted images. However, the training process of these methods requires an abundance of data to prevent overfitting, while there is no large enough database of authentically distorted images nowadays due to the expensive annotation process, leading to poor generalization [41, 74]. To tackle this, some researchers tried to explore more effective network architectures or training paradigms for BIQA tasks [37, 41, 46, 74]. For example, inspired by contrast learning,

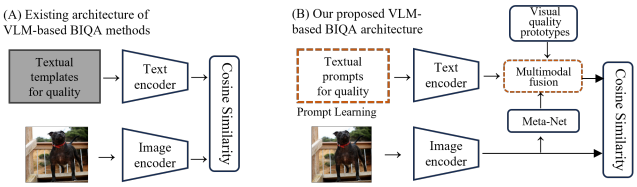


Figure 1: Existing VLM-based BIQA architecture vs. Ours.

several works tried to design unsupervised IQA networks for pre-training on large synthetically distorted databases or test-time adaptation [45], which aim to improve performances through extra unsupervised training or mitigating distribution shifts between training and test data. These studies focus on learning effective feature representations for quality perception via single visual modal information, while recent advances in Vision-Language Models (VLMs) have demonstrated the potential of textual information in building more generalizable feature representations [11, 42].

Therefore, as shown in Figure 1 (A), a few studies [65] have attempted to employ the contrastive language-image pre-training (CLIP) [42], a joint vision and language model trained with massive image-text pairs, to obtain the visual and textual embeddings to enhance the generalization of BIQA model. More recently, LIQE [81] improved the performance of the CLIP-based BIQA model by multi-task training. However, these methods still suffer from the following ***Challenges*** when applying VLMs to BIQA:

(1) These VLM-based methods rely solely on textual information for different quality-level representations (as shown in Figure 1 (A)), which neglect the valuable cross-modal features in multi-modal inputs. To enhance the quality perception, exploring methods for cross-modal fusion (as shown in Figure 1 (B)) beyond textual features is essential [87]. **(2)** Prompt engineering in VLMs-based IQA is extremely time-consuming. For example, CLIPIQA [65] devoted extensive efforts to find suitable prompts before training. Identifying the right prompt for the input text describing quality is a non-trivial task, as even slight changes in wording can significantly impact the performance and generalization (As shown in Figure 2, simply replacing the word *picture* in prompts with *photo* will reduce PLCC of the predicted MOSs against the ground-truth MOSs by 4.5%). Although CLIP-IQA+ [65] tried to finetune the prompts by backpropagation, the learned context is usually not generalizable enough [88]. **(3)** Due to the limited size of IQA datasets, effectively leveraging more datasets for training is essential, which requires the alignment of subjective quality scores across diverse datasets. Although several mixed-dataset training methods have been proposed [27, 57], they lack rigorous theoretical analysis and are tailored for single-modal IQA frameworks, neglecting text modules, thus they are not directly applicable to VLM-based BIQA models.

Therefore, in this paper, we propose a novel **M**ulti-**m**odal B**IQA** method with **P**rompt Learning, named MMP-IQA, which aims to achieve robust generalization by jointly training a quality-aware VTM on mixed datasets with fully utilizing cross-modal information. Inspired by [81], we first define five quality levels (including *bad*, *poor*, *fair*, *good*, and *perfect*) using an initialized textual template (such as *A photo with bad quality*) as a preparation before training. Then we leverage the pre-trained CLIP as the base architecture for

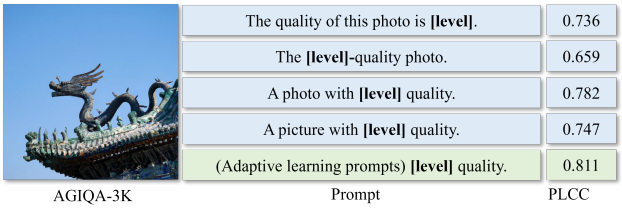


Figure 2: Example for *Challenge (2)*: prompts comparison *w.r.t.* PLCC results. Trained on Koniq-10k [18] and tested on AGIQA-3K [26] with our MMP-IQA. [level] is a quality level adjective.

VLM-based BIQA training, utilizing its exceptional generalization and zero-shot capabilities.

For ***Challenge (1)***, we propose to fuse the information from different modalities via cross-modal fusion. To be specific, utilizing labeled training samples, we compute average visual representations and text feature representations for each quality level using currently updated visual and text encoders respectively. Subsequently, a self-attention mechanism is employed to integrate cross-modal information conditioned on each sample, thereby enhancing textual representations and improving models' generalization capability.

Regarding ***Challenge (2)***, to avoid costly prompt engineering, inspired by recent prompt learning research in language models [6], we model the context words of quality prompts with learnable vectors, which are initialized with pre-trained word embeddings (as shown in Figure 3). The gradients of the learnable context vectors can be back-propagated through the text encoder, distilling the rich knowledge encoded in the parameters for image quality-relevant context. In addition, to prevent overfitting during prompt learning, we achieve the aforementioned cross-modal fusion by conditioning the encoded visual features of each image sample, which allows for the query embedding to be adjusted with context information. Consequently, MMP-IQA can efficiently capture more implicit generalizable elements that are vital for broader quality assessment.

For ***Challenge (3)***, we propose three quality evaluators for jointly training with a linearity-induced quality score, a relative quality score, and a dataset-specific absolute quality score. This strategy can align subjective quality scores across different datasets while preserving the unique characteristics of each dataset, leading to stable generalization. Besides, to emphasize the datasets and evaluators with larger loss values, we propose a dual automatic adjustment strategy for loss weights to adaptively balance the weights between different datasets and among various losses within the same dataset. The main contributions are summarized as follows:

- In this paper, we propose a novel Multi-modal BIQA method, named MMP-IQA. It can achieve robust generalization by jointly training a quality-aware CLIP on mixed datasets through the proposed conditional cross-modal fusion and prompt learning modules.
- An effective cross-modal fusion method is proposed for quality-aware VTM training, which can enhance textual representations and improve robustness by integrating cross-modal information conditioned on each sample.
- We propose an adaptive quality-aware prompt learning strategy, which can capture more generalizable elements without heavy prompt engineering during training.

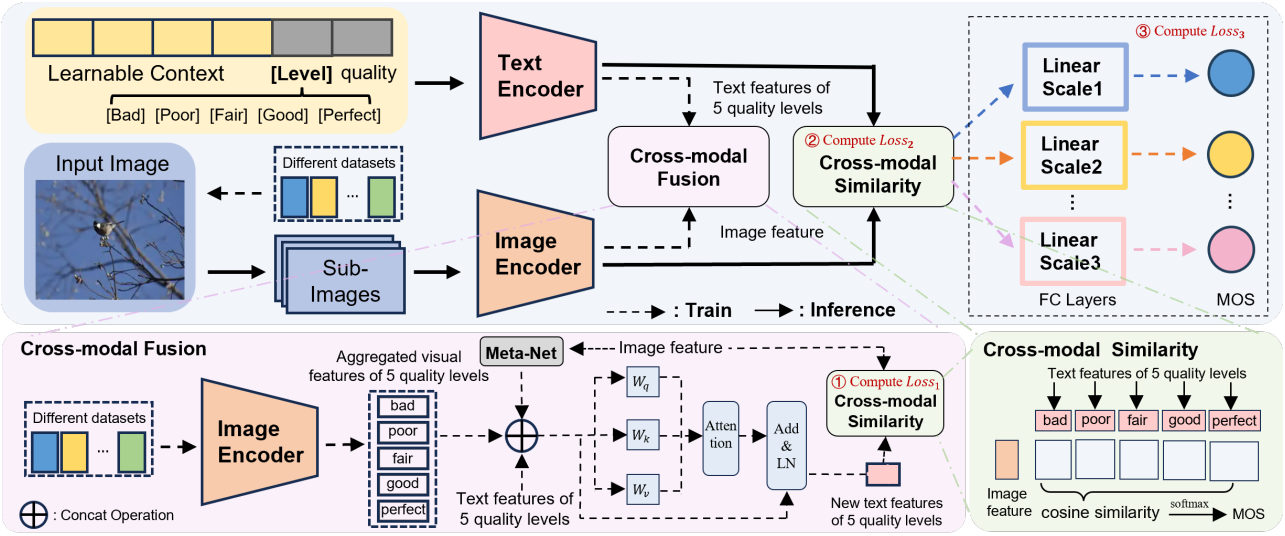


Figure 3: Overview of MMP-IQA. The upper part is the global picture of MMP-IQA. The lower light purple and green parts correspond to the Cross-modal Fusion module and Cross-modal Similarity module repectively. MMP-IQA consists of three stages: ① $Loss_1$: Relative quality assessor; ② $Loss_2$: Linearity-induced quality evaluator; ③ $Loss_3$: Dataset-specific absolute quality evaluator. The Learnable Context for prompt learning and the cross-modal fusion module are designed for *Challenge (2)* and *Challenge (1)* respectively. And ③ is jointly trained with ① and ② to address *Challenge (3)*. Meta-Net is a lightweight neural network with two FC layers.

- We provide the theoretical justification for the stability during mixed dataset training in MMP-IQA. Extensive experiments on diverse datasets, including authentic, synthetic, and AI-generated images, demonstrate MMP-IQA's superior performance and generalization capability.

*See Supp. Mat. for more literature reviews and related works on **Blind Image Quality Assessment (BIQA)**, **Prompt Learning**, **Mixed-datasets Training** and **CLIP Applications**.*

2 Methodology

2.1 Preliminaries and Overview

2.1.1 Problem Definition. Given an image $\boldsymbol{x} \in \mathbb{R}^D$ with potential distortions, the goal of BIQA model $f(\cdot)$ is to predict the perceptual quality of $\boldsymbol{x}$, aiming to match its MOS label $b(\boldsymbol{x}) \in \mathbb{R}$. This paper focuses on the CLIP-based BIQA model, while quality assessment is a regression problem. As such, we first discretize the quality levels to construct corresponding text descriptions. Inspired by the Likert-scale of five quality [16, 48], we set five quality levels from low to high, i.e. $\mathcal{C} = \{bad, poor, fair, good, perfect\}$, and each level corresponds to the standardized score $c \in \{1, 2, 3, 4, 5\}$. Then the quality score of $\boldsymbol{x}$ can be predicted by:

$$f(\boldsymbol{x}) = \sum\nolimits_{c=1}^{C} p(c \mid \boldsymbol{x}) \times c, \tag{1}$$

where C=5 is the number of quality levels, and $p(c \mid \boldsymbol{x})$ denotes the level probability that $\boldsymbol{x}$ belongs to quality level c. Through initializing the text template as $\mathcal{T}_{[\text{level}]}$ ($\mathcal{T}_c$): *A photo with* [level] *quality*, where [level] $\in \mathcal{C}$, the level probability $p(c \mid \boldsymbol{x})$ can be obtained by the pre-trained CLIP model.

2.1.2 Cross-modal Similarity. Specifically, the proposed quality-aware CLIP-based model $f(\cdot)$ includes a trainable image encoder $\boldsymbol{h}_\phi : \mathbb{R}^N \mapsto \mathbb{R}^K$ and another trainable text encoder $\boldsymbol{g}_\varphi : \mathcal{T}_{[\text{level}]} \mapsto \mathbb{R}^K$, the output dimensions of which are identical (i.e., $K = 512$ for CLIP). ϕ and φ are the trainable parameters of $\boldsymbol{h}_\phi$ and $\boldsymbol{g}_\varphi$ respectively. Informed by [81], for more accurate visual representation, we crop $U = 15$ sub-images to compute the aggregated visual embedding $\boldsymbol{H}(\boldsymbol{x}) \in \mathbb{R}^{U \times K}$ for each input image $\boldsymbol{x}$, since the image encoder $\boldsymbol{h}_\phi$ in CLIP requires input images to be cropped to a fixed size ($3 \times 224 \times 224$), which may affect the perceived quality of the images. Meanwhile, given the $C = 5$ textual descriptions $\mathcal{T}$ for quality levels, the aggregated textual embedding $\boldsymbol{G}(\mathcal{T}) \in \mathbb{R}^{C \times K}$ can be obtained by the lower-cased byte pair encoding (BPE) [47] and text encoder $\boldsymbol{g}_\varphi$. Then we can compute the cosine similarity between each image and textual quality description in an average manner:

$$\text{simi}\,(\boldsymbol{x}, \mathcal{T}_c) = \frac{1}{U} \sum_{u=1}^{U} \frac{\boldsymbol{H}_{u,:}(\boldsymbol{x}) \boldsymbol{G}_{c,:}^{\mathsf{T}}(\mathcal{T})}{\left\|\boldsymbol{H}_{u,:}(\boldsymbol{x})\right\|_2 \left\|\boldsymbol{G}_{c,:}(\mathcal{T})\right\|_2}, \tag{2}$$

where $\boldsymbol{H}_{u,:}$ indexes the u-th row vector in the matrix, and $\bullet^{\mathsf{T}}$ denotes the transpose of a matrix. After matching the image $\boldsymbol{x}$ to all textual quality descriptions, we can obtain each quality level probability for $\boldsymbol{x}$ by softmax:

$$p(c \mid \boldsymbol{x}) = \frac{\exp\left(\text{simi}(\boldsymbol{x}, \mathcal{T}_c)/\tau_1\right)}{\sum_c \exp\left(\text{simi}(\boldsymbol{x}, \mathcal{T}_c)/\tau_1\right)}, \tag{3}$$

where $c = 1, 2, \ldots, 5$, corresponding to 5 different quality levels. τ_1 is a learnable hyper-parameter.

2.1.3 Overview. Figure 3 shows the overview of our MMP-IQA, which consists of three stages: ① Relative quality assessor; ② Linearity-induced quality evaluator; ③ Dataset-specific absolute quality evaluator.During the mixed datasets training, the three evaluators are jointly optimized to address *Challenge **(3)***. For *Challenges **(1)** and **(2)***, we propose conditional cross-modal fusion and prompt learning modules respectively to enhance the generalization capability without prompt engineering. **Each part is detailed below**.

2.2 Adaptive Prompt Learning

For *Challenge (2)*, in Figure 3, we propose an adaptive prompt learning method for quality-level description, eliminating the expensive manual prompt tuning by modeling context words with continuous vectors that are end-to-end learned from raw IQA datasets. Specifically, the prompt input to the text encoder $\boldsymbol{g}_\varphi$ is structured as:

$$\mathcal{T}_{[\text{level}]} = [\text{V}]_1[\text{V}]_2 \dots [\text{V}]_M[\text{level}] \text{ quality}, \tag{4}$$

where M is the number of trainable context tokens, and each $[\text{V}]_m$ ($m \in \{1, \dots, M\}$) in Eq. (4) is a learnable vector with the same dimension K as word embeddings. In our method, we initialize the learnable context with the "textual template": *A photo with* [level] *quality*. Thus $M = 3$ in Eq. (4). Inspired by [81], we set [level]$\in C = \{$*bad*, *poor*, *fair*, *good*, *perfect*$\}$. After training until convergence, we can obtain the quality level probability by Eq. (3).

To investigate the impact of M value on model performance, instead of placing the token of "[Level] quality" at the end of the sequence in Eq. (4), we can also position it in the middle with $[\text{V}]_4 \dots [\text{V}]_M$ randomly initialized:

$$\mathcal{T}_{[\text{level}]} = [\text{V}]_1[\text{V}]_2[\text{V}]_3[\text{level}] \text{ quality } [\text{V}]_4 \dots [\text{V}]_M. \tag{5}$$

The experimental results are shown in the Section 3.4.1.

2.3 Cross-modal fusion

For *Challenge (1)*, we propose a conditional cross-modal fusion module to improve generalization by leveraging multi-modal information. The core idea is to construct the cross-modal context for a more robust quality-level probability calculation, which is conditioned on each input image. The module includes three main steps: **(i)** Calculating visual quality prototype; **(ii)** Constructing cross-modal textual representations; **(iii)** Computing quality level probability.

2.3.1 Calculating Visual Quality Prototype. Recent studies have identified an inherent domain gap [29] between visual and textual embeddings in VLMs, causing them to form separate clusters in the embedding space and hindering effective pair-wise mapping. To address this, inspired by the concept of visual prototypes [52] in the domain of image classification, where the average feature of each class serves as a prototype, we define the visual quality prototype as the shared characteristics of each quality level. Specifically, we **first** normalize the MOS labels of each dataset to the range of [1, 5] to align with the standardized score from the textual module in Eq. (1). This is reasonable in MMP-IQA, since the loss for each dataset is calculated separately in subsequent training, and each dataset's unique characteristics are preserved. To retain each dataset's unique characteristics during subsequent training, we calculate each dataset's loss separately and propose three quality assessors for joint training to align label ranges across datasets. Details on mixed-dataset training are in Section 2.4. **Then** we evenly divide them into five intervals (i.e. [1.0, 1.8), [1.8, 2.6),..., [4.2, 5.0]), corresponding to five quality levels: *bad, poor, fair, good*, and *perfect*. For these datasets, samples with the same quality level are grouped together, forming sets like $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_5$. On this basis, we define the visual quality prototype of quality level c as:

$$\mathcal{P}_c = \frac{1}{|\mathcal{D}_c|} \sum_{i=1}^{|\mathcal{D}_c|} \boldsymbol{h}_\phi\left(\boldsymbol{x}_{c,i}\right), \boldsymbol{x}_{c,i} \in \mathcal{D}_c, c = 1, \dots, 5. \tag{6}$$

They are computed through a forward pass in the image encoder $\boldsymbol{h}_\phi$ at the start of each training epoch. Note that these visual quality prototypes will continuously update after the image encoder $\boldsymbol{h}_\phi$ updates during training, and they focus not on specific distortion types but rather on overall quality levels, representing the key features of each quality level. Thus these visual quality prototypes can be used as the visual context to refine the embeddings.

2.3.2 Cross-modal Textual Representations. This module consists of two main steps: (1) multi-modal concatenation and (2) instance-conditioned multi-modal fusion. Specifically, in the **Step (1)**, given the text templates $\mathcal{T}$, we define $\mathcal{W}_c = \boldsymbol{G}_{c,:}(\mathcal{T})$ as the textual embedding of quality level c ($c = 1, \dots, 5$). Based on the calculated visual quality prototypes in Eq. (6), as the preparation before cross-modal fusion, we augment the textual context with the visual information. That is **Context** $= [\mathcal{W}, \mathcal{P}]$ with $\mathcal{P} = [\mathcal{P}_1, \dots, \mathcal{P}_5] \in \mathbb{R}^{5\times K}$ and $\mathcal{W} = [\mathcal{W}_1, \dots, \mathcal{W}_5] \in \mathbb{R}^{5\times K}$, which is the concatenation of all visual and textual representations for each quality level. Integrating prototypes from multiple modalities enables the quality-aware VLM model to adapt and fuse cross-modal information, surpassing the basic visual-textual matching.

Besides, in the **Step (2)**, since the image quality is also related to the distortion types and scene contents [85], we argue that instance-conditioned context can generalize better, which can shift the focus away from a specific set of quality levels to each input image instance $\boldsymbol{x}$, and hence to the entire IQA task. Based on this insight, we propose conditional cross-modal fusion to improve the robustness of the proposed prompt learning strategy. Specifically, we denote the adaptation function for cross-modal fusion as $\mathcal{R}(\cdot)$, which takes in the query instance $\boldsymbol{z} = \boldsymbol{h}_\phi(\boldsymbol{x})$ and context information as bags and outputs the set of adjusted embeddings while maintaining the permutation invariant. That is:

$$[\tilde{z}, \tilde{\textbf{Context}}] = \mathcal{R}\left([\mathcal{S}(\boldsymbol{z}), \textbf{Context}]\right), \tag{7}$$

where $\tilde{\textbf{Context}} = [\tilde{\mathcal{W}}, \tilde{\mathcal{P}}]$ is instance-conditioned context after cross-modal fusion, and $\mathcal{S}(\cdot)$ denotes a lightweight neural network named Meta-Net, which is used to learn the implicit high-level feature about scene and distortion type from the input image instance $\boldsymbol{x}$ after encoding by $\boldsymbol{z} = \boldsymbol{h}_\phi(\boldsymbol{x})$. According to Eq. (7) and Figure 1 (B), the features extracted by Meta-Net enable the context to become instance-conditioned context through the cross-modal fusion module. The Meta-Net includes two FC layers, maintaining consistent input and output dimensions. More details of Meta-Net are described in Supp. Mat. In our method, we use the self-attention (SA) mechanism [31, 62] as the cross-modal fusion function $\mathcal{R}(\cdot)$. Due to the permutation invariance, SA excels at producing adapted embeddings, even with long dependencies, making it well-suited for the adaptation function. See Supp. Mat. for SA details. Finally, we can obtain cross-modal textual representations $\tilde{\boldsymbol{G}}(\mathcal{T})$ for each quality level with a residual structure:

$$\tilde{\boldsymbol{G}}(\mathcal{T}) = \tilde{\mathcal{W}} + \boldsymbol{G}(\mathcal{T}). \tag{8}$$

2.3.3 Computing Quality Level Probability. By substituting the cross-modal textual representations $\tilde{\boldsymbol{G}}(\mathcal{T})$ into Eq. (2) and Eq. (3), we can calculate a new quality level probability $\tilde{p}(c \mid \boldsymbol{x}), (c =$

$1, \ldots, 5$) for each image sample $\boldsymbol{x}$:

$$\tilde{\text{simi}}\,(\boldsymbol{x}, \mathcal{T}_c) = \frac{1}{U} \sum_{u=1}^{U} \frac{\boldsymbol{H}_{u,:}(\boldsymbol{x}) \tilde{\boldsymbol{G}}_{c,:}^{\mathsf{T}}(\mathcal{T})}{\left\|\boldsymbol{H}_{u,:}(\boldsymbol{x})\right\|_2 \left\|\tilde{\boldsymbol{G}}_{c,:}(\mathcal{T})\right\|_2}, \tag{9}$$

$$\tilde{p}(c \mid \boldsymbol{x}) = \frac{\exp\left(\tilde{\text{simi}}(\boldsymbol{x}, \mathcal{T}_c)/\tau_2\right)}{\sum_c \exp\left(\tilde{\text{simi}}(\boldsymbol{x}, \mathcal{T}_c)/\tau_2\right)}, \tag{10}$$

where τ_2 is a learnable hyper-parameter, independent of τ_1.

2.4 Mixed-datasets Training

For *Challenge (3)*, we propose three quality assessors for joint training that align label ranges across datasets, preserve each dataset's unique characteristics, and account for linear differences between predicted and actual labels. They are balanced by a dual automatic loss weight adjustment strategy.

2.4.1 Relative Quality Assessor. To align label ranges across datasets, we consider the pairwise learning-to-rank model estimation for BIQA. Specifically, for an image pair $(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d})$ from the same d-th dataset, we compute a binary label by their ground-truth MOSs $b(\boldsymbol{x}_{1,d})$ and $b(\boldsymbol{x}_{2,d})$:

$$q(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d}) = \begin{cases} 1 & \text{if } b(\boldsymbol{x}_{1,d}) \geq b(\boldsymbol{x}_{2,d}) \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

After predicting the relative quality score based on Eqs. (3) and (1), we estimate the probability of $\boldsymbol{x}x_{1,d}$ perceived better than $\boldsymbol{x}_2^d$ under the Thurstone's model [59]:

$$\hat{q}(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d}) = \Phi\left(\frac{\boldsymbol{f}(\boldsymbol{x}_{1,d}) - \boldsymbol{f}(\boldsymbol{x}_{2,d})}{\sqrt{2}}\right), \tag{12}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function with the variance fixed to one. We use the fidelity loss [61] as the measure of statistical distance:

$$\begin{aligned} L_1^{(d)}(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d}) = 1 - \sqrt{q(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d})\hat{q}(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d})} \\ - \sqrt{(1 - q(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d}))(1 - \hat{q}(\boldsymbol{x}_{1,d}, \boldsymbol{x}_{2,d}))} \end{aligned} \tag{13}$$

where $L_1^{(d)}$ means the fidelity loss on the d-th dataset.

2.4.2 Linearity-induced Quality Assessor. Through substituting Eq. (10) into Eq. (1) in the cross-modal fusion module, we can predict the quality score $Q_{i,d} = \tilde{\boldsymbol{f}}(\boldsymbol{x}_{i,d})$ for the i-th image $\boldsymbol{x}_{i,d}$ in the d-th dataset. To maximize the prediction linearity between perceptual and subjective quality scores, we define a linearity-induced loss using Pearson's linear correlation coefficient (PLCC), which is differentiable and effectively measures linearity:

$$\text{PLCC}_d = \frac{\sum_i \left(Q_{i,d} - \bar{Q}_{p,d}\right)\left(b(\boldsymbol{x}_{i,d}) - \bar{b}_{g,d}\right)}{\sqrt{\sum_i \left(Q_{i,d} - \bar{Q}_{p,d}\right)^2 \sum_i \left(b(\boldsymbol{x}_{i,d}) - \bar{b}_{g,d}\right)^2}}, \tag{14}$$

$$L_2^{(d)} = \left(1 - \text{PLCC}_d\right)/2, \tag{15}$$

where $Q_{i,d}$ and $b(\boldsymbol{x}_{i,d})$ denote the predicted MOS score and actual MOS score for image sample $\boldsymbol{x}_{i,d}$ respectively. $\bar{Q}_{p,d}$ and $\bar{b}_{g,d}$ are the averages of the predicted and actual MOS scores respectively in the d-th IQA dataset.

2.4.3 Absolute Quality Evaluator. To preserve each dataset's unique characteristics while aligning label ranges across different datasets, for the i-th image $\boldsymbol{x}_{i,d}$ in the d-th IQA dataset, we propose to compute the dataset-specific absolute quality score $Y_{i,d}$ by:

$$Y_{i,d} = \boldsymbol{F}_d\left(\boldsymbol{f}(\boldsymbol{x}_{i,d})\right), \tag{16}$$

where $\boldsymbol{F}_d(\cdot)$ is the FC layer for linearly scaling the MOS on the d-th IQA dataset, and $\boldsymbol{f}(\boldsymbol{x}_{i,d})$ is the predicted quality score based on Eqs. (10) and (1). Then we can define the absolute quality loss for the d-th dataset by MAE (ℓ_1) loss:

$$L_3^{(d)} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left|Y_{i,d} - b(\boldsymbol{x}_{i,d})\right|, \tag{17}$$

where N_d is the number of images in the d-th dataset.

2.4.4 Dual Automatic Weight Adjustment. Based on Eqs. (13) (15) and (17), we define the final loss function as a linearly weighted summation of the three individual losses:

$$L_{total} = \lambda_1(t)L_1 + \lambda_2(t)L_2 + \lambda_3(t)L_3, \tag{18}$$

$$L_k = \sum\nolimits_{d=1}^{D} w_k^{(d)} L_k^{(d)}, \ (k = 1, 2, 3), \tag{19}$$

where D is the number of datasets. The loss weights at the t-th iteration are automatically computed through the proposed dual automatic adjustment strategy:

(1) **Inner Automatic Weight**: each $w_k^{(d)}$ in Eq. (19) is a softmax-weighted average of losses over all datasets, i.e. $w_k^{(d)} = e^{L_k^{(d)}} / \sum_d^D e^{L_k^{(d)}}$.

(2) **Outer Automatic Weight**: we update $\lambda_k(t)(k = 1, 2, 3)$ in Eq. (18) according to the relative descending rate [32], that is:

$$\lambda_k(t) = \frac{\exp\left(\beta_k(t-1)/\tau_3\right)}{\sum_{j=1}^{3} \exp\left(\beta_j(t-1)/\tau_3\right)}, \beta_k(t-1) = \frac{L_k(t-1)}{L_k(t-2)}, \tag{20}$$

where τ_3 is a hyper-parameter fixed as 1.

Based on Eqs. (18)-(20) during mixed-datasets training, we have:

THEOREM 2.1 (GENERALIZATION BOUND). *If the generalization error of each $L_k^{(d)}$ decays as $O(1/\sqrt{N_d})$ (N_d being the sample size of $\mathcal{D}_d$), then the upper bound of the generalization error G_e is:*

$$G_e = \mathbb{E}[L_{total}] - L_{total} \leq O\left(\sum\nolimits_{d=1}^{D} \frac{1}{\sqrt{N_d}} \left(\sum\nolimits_{k=1}^{3} \lambda_k w_k^{(d)}\right)\right), \tag{21}$$

where $\mathbb{E}[L_{total}]$ is the expected loss, and L_{total} is the empirical loss.

From the theoretical result Eq. (21) in Theorem 2.1, we can conclude that the generalization error is primarily governed by the smallest dataset (since $1/\sqrt{(N_d)}$ decays as N_d increases). In addition, Theorem 2.1 can provide the theoretical justification for the stability during joint training across multiple datasets. The detailed analysis and proof of Theorem 2.1 are provided in the Supp. Mat.

3 Experiments

3.1 Experimental Settings

We jointly train our method on the six IQA datasets: LIVE [48], CSIQ [24], and KADID-10k [30], which contain synthetic distortions, and LIVE Challenge [16] (referred to as CLIVE in Table 1), BID [7], and KonIQ-10k [18], which feature realistic distortions. Details of these datasets can be found in Supp. Mat. For each dataset,

Table 1: Main results on six datasets. The best (second-best) results are highlighted in bold (with an underline).

Dataset	KADID-10k [30]		LIVE [48]		CSIQ [24]		BID [7]		CLIVE [16]		KonIQ-10k [18]	
Criterion	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [38]	0.389	0.442	0.908	0.904	0.631	0.719	0.573	0.618	0.446	0.507	0.415	0.438
ILNIQE [76]	0.565	0.611	0.887	0.894	0.808	0.851	0.548	0.494	0.509	0.518	0.894	0.534
Ma19 [36]	0.465	0.501	0.922	0.923	0.926	0.929	0.373	0.399	0.336	0.405	0.360	0.398
PaQ2PiQ [72]	0.403	0.448	0.544	0.558	0.697	0.766	0.719	0.700	0.732	0.755	0.722	0.716
KonCept [18]	0.503	0.515	0.673	0.619	0.631	0.645	0.816	0.825	0.778	0.799	0.911	0.924
MUSIQ [21]	0.572	0.584	0.837	0.818	0.697	0.766	0.744	0.774	0.785	0.828	0.915	**0.937**
Causal-IQA [85]	0.838	0.878	0.937	0.927	0.914	0.915	0.862	0.885	0.822	0.896	0.913	0.917
HyperIQA [54]	0.872	0.869	0.966	**0.968**	0.934	**0.946**	0.848	0.868	0.855	0.878	0.900	0.915
TReS [17]	0.881	0.879	0.965	0.963	0.902	0.923	0.853	0.871	0.846	0.877	0.907	0.924
UNIQUE [79]	0.884	0.885	0.961	0.952	0.902	0.921	0.852	0.875	0.854	0.884	0.895	0.900
LIQE [81]	0.930	**0.931**	0.970	0.951	0.936	0.939	0.875	0.900	0.904	0.910	0.919	0.908
MMP-IQA	**0.939**	0.897	**0.973**	0.948	**0.946**	0.935	**0.887**	**0.913**	**0.916**	**0.921**	**0.926**	0.910

Table 2: SRCC results on the three IQA datasets under the cross-dataset setup. The subscript "r" stands for models trained on KonIQ-10K. The best results are boldfaced.

Dataset	TID2013 [40]	SPAQ [14]	PIPAL [20]
NIQE [38]	0.314	0.578	0.153
PaQ2PiQ [72]	0.423	0.823	0.400
MUSIQ$_r$ [21]	0.584	0.853	0.450
UNIQUE [79]	0.768	0.838	0.444
LIQE [81]	0.811	0.881	0.478
MMP-IQA	**0.824**	**0.887**	**0.490**

70% of images are randomly selected for training and 10% for validation, leaving the remaining 20% for testing. We use ViT-B/32 [42] as the visual encoder and GPT-2 [44] with 63M parameters as the text encoder. The model is trained by minimizing the objective in Eq. (18) using AdamW [34] with a weight decay of 10^{-3}. The initial learning rate is 5×10^{-6}, adjusted by a cosine annealing schedule [33]. We optimize MMP-IQA for 100 epochs with a batch size of 16 on the LIVE, CSIQ, BID, and CLIVE datasets, and 64 on KonIQ-10k and KADID-10k. During training and inference, we randomly crop 3 and 15 sub-images at 224×224×3 from the original images while maintaining aspect ratios. All experiments are run on a single NVIDIA Tesla A100 GPU. We evaluate MMP-IQA by two typical IQA metrics: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC) [4].

3.2 Performances Comparisons

3.2.1 Main Results. We compare the PLCC and SRCC results of the proposed MMP-IQA against three opinion-unaware BIQA models, including NIQE [38], ILNIQE [76], and Ma19 [36], and eight data-driven DL-based methods, including PaQ2PiQ [72], KonCept [18], Causal-IQA [85], MUSIQ [21], HyperIQA [54], TreS [17], UNIQUE [79] and LIQE [81], where LIQE is the SOTA CLIP-based BIQA model. Following [81], the publicly available implementations are either directly adopted or re-trained on our datasets with the training codes provided by the respective authors for competing models. HyperIQA, Causal-IQA and TreS are separately trained on each dataset. Equipped with more sophisticated backbone networks and trained on KonIQ-10k, KonCept and MUSIQ deliver better performance. UNIQUE, LIQE and MMP-IQA are jointly trained on six datasets. From Table 1 we can observe that MMP-IQA performs best overall, demonstrating its superiority. Specifically, **(1)** Opinion-unaware methods (NIQE, ILNIQE, Ma19) and PaQ2PiQ demonstrate suboptimal performances (detailed analysis in Supp. Mat.) **(2)** Performances of KonCept and MUSIQ: With advanced backbone networks and training on KonIQ-10k, both of them deliver better performance on KonIQ-10k but struggle with generalization on other datasets, particularly with complex synthetic distortions in KADID-10k. Demonstrating the challenges in handling cross-distortion scenarios. **(3)** For HyperIQA, Causal-IQA and TreS: These models are separately trained on each dataset with distinct parameters, achieving strong performances across six datasets. **(4)** Strengths of UNIQUE, LIQE and MMP-IQA: The pairwise learning-to-rank training strategy enables UNIQUE and LIQE to learn from multiple datasets simultaneously with a single set of weights. The overall performances of MMP-IQA are notably better than that of UNIQUE and LIQE on both authentic and synthetic IQA datasets, verifying the effectiveness of the proposed method. Although MMP-IQA achieved the best performance on the BID dataset, its SRCC score was only 0.887. This may be due to the limited variety of distortions it contains (only blur), whereas MMP-IQA is better suited for handling more complex distortions. Although MMP-IQA achieved second-best or suboptimal PLCC results on KADID-10k, LIVE, CSIQ, and KonIQ-10k, it excelled in SRCC performances. This is likely due to the Relative Quality Assessor (i.e., Eq. (13)) during the mixed-dataset training, which enhances the model's ability to better capture the relative ranking of image quality rather than the absolute quality scores, improving generalization across diverse datasets.

For a comprehensive comparison of LIQE and MMP-IQA regarding algorithmic differences, please refer to Supp. Mat. The two models demonstrate similar parameter counts (LIQE: 151.3M **vs.** MMP-IQA: 152.5M) and comparable training speeds (LIQE: 42.6 *samples / sec* **vs.** MMP-IQA: 40.2 *samples / sec*) on a single NVIDIA A10 GPU, confirming their analogous computational scales.

3.2.2 Cross-Dataset Evaluation. We compare the generalizability of MMP-IQA against competitive BIQA models in a more challenging cross-dataset setting. Specifically, we test models trained in Table 1 on the full TID2013 [40] and SPAQ [14], which contain synthetic and realistic camera distortions, respectively. Besides, we

Table 3: Ablation study: SRCC results of MMP-IQA variants. The best (second-best) results are highlighted in bold (with underlines).

Variants of **MMP-IQA**	LIVE [48]	KADID-10k [30]	KonIQ-10k [18]	CSIQ [24]	BID [7]	CLIVE [16]	Mean
(1) Pretrained CLIP ($C = 2$)	0.741	0.732	0.684	0.783	0.629	0.630	0.699
(2) Fine-Tuned CLIP ($C = 2$)	0.928	0.891	0.904	0.897	0.834	0.872	0.887
(3) Frozen Textual Encoder	0.941	0.903	0.901	0.911	0.842	0.885	0.897
(4) Fixed Textual Prompts	0.962	0.928	0.923	0.928	0.871	0.909	0.920
(5) Without Cross-modal Fusion	0.955	0.925	0.916	0.910	0.858	0.887	0.909
(6) Equal Inner Loss Weights	0.965	**0.942**	0.925	0.934	0.867	0.905	0.923
(7) Equal Outer Loss Weights	0.968	0.936	0.921	**0.946**	0.883	0.893	0.925
(8) Without Meta-Net	0.971	0.934	0.924	0.941	0.880	0.912	**0.927**
Full **MMP-IQA**	**0.973**	0.939	**0.926**	**0.946**	**0.887**	**0.916**	**0.931**

Table 4: PLCC and SRCC results of MMP-IQA trained with different loss combinations. *ALL* means the full MMP-IQA.

Loss	L_1	L_2	L_1+L_2	L_1+L_3	L_2+L_3	*ALL*
PLCC	0.905	0.910	0.912	0.916	0.914	**0.921**
SRCC	0.917	0.909	0.916	0.923	0.926	**0.931**

also employ images from the training set of PIPAL [20] as a test set, the MOSs of which are publicly available. The results in Table 2 can illustrate the robustness and generalization of the proposed method.

3.3 Ablation Studies

In Tables 3 and 4, we conduct ablation studies to verify the design rationality of MMP-IQA following the same experimental setups as Table 1. The variants of MMP-IQA include: **(1)** Implementing a pre-trained CLIP baseline which adopts a Likert-scale of two quality levels [65], i.e., $c \in C = \{bad, good\} = \{1, 2\}$. **(2)** Fine-tuning (1) on IQA datasets. **(3)** Freezing the text encoder $\boldsymbol{g}_\varphi$ during training. **(4)** Fixing textual prompts as *A photo with* [level] *quality* without prompt learning. **(5)** Implementing MMP-IQA without cross-modal fusion. **(6)** Adopting equal inner loss weights, i.e., the weights $w_k^{(d)}$ is same set to 1 in Eq. (19). **(7)** Adopting equal outer loss weights, i.e., the weights $\lambda_k(t)$ is same set to 1 in Eq. (18). **(8)** Implementing conditional cross-modal fusion without Meta-Net.

From Table 3, we can observe that each component designed for three Challenges is effective in MMP-IQA. Specifically, **(1)** The pre-trained CLIP is not good at BIQA, whose performance can be improved by fine-tuning on IQA datasets. **(2)** The MMP-IQA model with 5 quality levels in the textual template performs better than that with only 2 quality levels. **(3)** Performance is negatively affected when the language encoder is frozen, which may be because quality-related concepts have not been sufficiently captured during the pre-training stage of CLIP. **(4)** MMP-IQA works better with the trainable textual prompts than with fixed textual prompts, revealing the validity of the proposed adaptive prompt learning method for *Challenge (2)*. **(5)** The full MMP-IQA method performs favorably against the variant trained without the cross-modal fusion information, verifying the effectiveness of the proposed cross-modal fusion module for *Challenge (1)*. **(6)** It is worth noting that the dual automatic weight adjustment scheme can not only liberate us from laborious hyper-parameter tuning but also achieve improvements over adopting equal task weightings, verifying the effectiveness of the proposed cross-modal fusion module for *Challenge (3)*. **(7)** The full MMP-IQA method using a Meta-Net for constructing the cross-modal textual representations performs better than that without Meta-Net shows that the instance-conditioned context can generalize better as it captures more implicit elements related to image quality. **(8)** The 6-th variant of MMP-IQA (Equal Inner Loss Weights) achieves the best results on KADID-10k, likely because KADID-10k has the enough samples in the mixed dataset, and fixed equal inner loss weights for each dataset help prevent excess weight allocation to smaller datasets. The 7-th variant of MMP-IQA (Equal Outer Loss Weights) performs best on CSIQ, likely due to its smaller sample size (fewer than 1000). Fixed equal outer loss weights for each assessor are advantageous for CSIQ, as they effectively retain CSIQ-specific characteristics within the Absolute Quality Evaluator.

In addition, we investigate the impact of each quality assessor on the performance of MMP-IQA by training MMP-IQA with different combinations of losses and summarize the mean SRCC and mean PLCC over six datasets in Table 4, where *ALL* means MMP-IQA trained with all losses. From Table 4, we can observe that the full MMP-IQA model jointly trained with three quality assessors performs better than the model trained with only one or two quality assessors, which proves the superiority of the proposed dual mixed datasets training strategy. Concretely speaking: **(1)** Whether trained the Cross-modal Similarity module based on L_1 or L_2, including the task of absolute quality evaluation for each dataset is beneficial for BIQA, reflecting the importance of preserving the unique characteristics of each IQA dataset during mixed dataset training. **(2)** The full MMP-IQA including the three quality evaluators together performs best, which reflects the cooperative relationship among the relative quality assessor, linearity-induced quality evaluator, and the dataset-specific absolute quality evaluator.

3.4 Additional Experiments

3.4.1 The Impact of Learnable Context Length. By adjusting the M value in Eq. (5), we can investigate the impact of learnable context length in the proposed adaptive prompt learning module on the performances of MMP-IQA. We conduct these experiments following the same settings as Table 1, the results on the CLIVE and BID datasets are recorded in Figure 4, from which we can observe that the impact of M on model performance is slight. Therefore, we set $M = 3$ in the MMP-IQA for relatively stable training.

3.4.2 The Impact of Different Initialization Modes. As described in Section 2.2, we manually initialize the learnable tokens in Eq.(4) with the phrase "A photo with," which aligns with natural human expression. We investigate the impact of different initialization methods and templates, where random initialization is generated from a zero-mean Gaussian distribution with a standard deviation of

Table 5: Comparison between random and manual initializations on learnable tokens (highlighted in green) of different templates. "Initialize randomly" means the learnable tokens in Eq. (4) are randomly initialized with the learnable context length $M = 3$.

Initialization mode	PLCC	SRCC
"The quality of this photo is [level]"	0.919	0.926
"The [level]-quality photo"	0.917	0.930
"A picture with [level] quality"	0.920	**0.931**
"A photo with [level] quality"	**0.921**	**0.931**
Initialize randomly	0.919	0.929

Table 6: Zero-shot results (SRCC) on the AIGC datasets. The best (second-best) results are in bold (with underlines).

Models	AGIQA-3K [26]	AGIQA-1K [82]	AIGCIQA2023 [67]
HyperIQA [54]	0.629	0.525	0.721
TReS [17]	0.646	0.510	0.741
UNIQUE [79]	0.666	0.460	**0.761**
MUSIQ [21]	0.629	0.525	0.736
PaQ2PiQ [72]	0.502	0.538	0.643
CLIPIQA [65]	0.658	0.341	0.609
CLIPIQA+ [65]	0.683	0.446	0.716
MANIQA [71]	0.695	**0.618**	0.728
LIQE [81]	**0.721**	0.579	0.744
MMP-IQA	**0.745**	**0.586**	**0.767**

0.02. For a fair comparison with the initialization mode ("A photo with [level] quality") used in MMP-IQA, the learnable context length M is set to 3 in "Initialize randomly". Table 5 summarizes the mean SRCC and mean PLCC across six datasets with the same experimental setups as Table 1, from Table 5 we can observe that manual initialization does not result in a significant performance improvement. This is likely because the adaptive prompt learning process is a continuous optimization in the word embedding space, demonstrating the robustness of MMP-IQA to different prompt initializations.

3.4.3 Cross-database Results on AIGC Datasets. To further verify the effectiveness of MMP-IQA, we train MMP-IQA on the single IQA dataset KonIQ-10k and test on three AIGC datasets without fine-tuning, including AGIQA-1K [82], AGIQA-3K [26] and AIGC2023 [67]. We add two new methods for comparison, including CLIPIQA [65], CLIPIQA+ [65] (the variant of CLIPIQA), and MANIQA [71]. The SRCC results in Table 6 prove the advanced generalization capability of our MMP-IQA.

3.4.4 Change in the Distribution of Visual Quality Features Before and After Training. As stated in Section 2.3.1 in the main text, after computing the visual quality prototypes by the image encoder $\boldsymbol{h}_\phi$ before training, these visual quality prototypes will be continuously updated after the image encoder $\boldsymbol{h}_\phi$ updates during training. To validate the effectiveness of the proposed visual quality prototypes, we randomly selected 1000 samples from each of the five quality levels both before training and after 80 epochs of training. We then applied t-SNE to reduce the dimensionality of their visual quality features and plotted the resulting scatter plots, as shown in Figure 5. From Figure 5, we can observe that, after training, the samples within the same quality level are more concentrated compared to before training. This not only demonstrates that the image encoder $\boldsymbol{h}_\phi$ in MMP-IQA has been effectively trained, but also

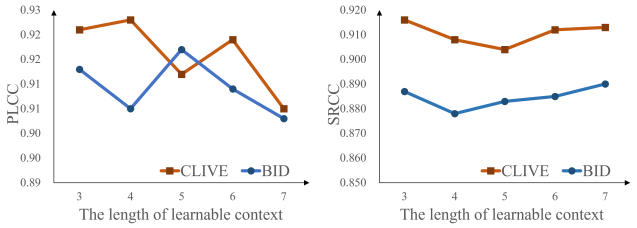
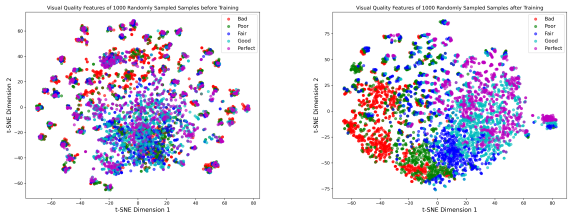


Figure 4: The impact of the length of learnable context on the performance of MMP-IQA in adaptive prompt learning.



(a) Distribution of visual quality features before training. **(b) Distribution of visual quality features after training.**

Figure 5: The difference between the distribution of visual quality features before training and that after 80 epochs of training.

confirms that the visual quality prototypes proposed in Section 2.3.1 have become increasingly robust and effective. Specifically, they focus not on specific distortion types but rather on overall quality levels, representing the key features of each quality level. Therefore, these visual quality prototypes can be used as visual context to refine the embeddings, which can improve the performance of MMP-IQA.

3.4.5 More Experimental Results. We also conducted more experiments to further explore several aspects, including (1) the importance of the token of "quality" in the template, (2) The impact of different mixed-dataset training methods on MMP-IQA, and (3) some visual examples predicted by MMP-IQA. For detailed results and discussion, please refer to Supp. Mat.

4 Conclusion

To address the generalization issues in BIQA and three challenges associated with existing CLIP-based BIQA methods, in this work, we propose a novel multi-modal BIQA approach with prompt learning, named MMP-IQA, which jointly trained a quality-aware VTM on mixed datasets with cross-modal fusion. MMP-IQA has three significant capabilities: **(1)** Enhanced Generalization Capacity due to multi-modal information fusion; **(2)** Adaptive Quality-aware Prompting Capacity for capturing more generalizable elements; **(3)** Robust and Parameter-free Mixed-training Capability addressing the issue of small-scale IQA datasets during the multi-modal BIQA training process. Extensive experiments verify the superiority of MMP-IQA.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under contract 62088102. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. 2024. Quality-aware image-text alignment for real-world image quality assessment. *arXiv preprint arXiv:2403.11176* 5, 6 (2024).

[2] Peter L Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3, Nov (2002), 463–482.

[3] Vidmantas Bentkus. 2004. On Hoeffding's inequalities. (2004).

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Qizhou Chen, Taolin Zhang, Dongyang Li, Longtao Huang, Hui Xue, Chengyu Wang, and Xiaofeng He. 2024. Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning. *arXiv preprint arXiv:2405.03279* (2024).

[7] Alexandre Ciancio, Eduardo AB da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al. 2010. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing* 20, 1 (2010), 64–75.

[8] Hugo Costa, Giles M Foody, and Doreen S Boyd. 2017. Using mixed objects in the training of object-based image classifications. *Remote Sensing of Environment* 190 (2017), 188–197.

[9] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-Based Named Entity Recognition Using BART. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*. Association for Computational Linguistics, 1835–1845.

[10] Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. 2023. IRNeXt: Rethinking Convolutional Network Design for Image Restoration. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 6545–6564.

[11] Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. 2023. Multimodal parameter-efficient few-shot class incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3393–3403.

[12] Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. 2022. Image generation: A review. *Neural Processing Letters* 54, 5 (2022), 4609–4646.

[13] Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)* (1978).

[14] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3677–3686.

[15] Yingjie Feng, Sumei Li, and Sihan Hao. 2021. An Error Self-learning Semi-supervised Method for No-reference Image Quality Assessment. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 1–5.

[16] Deepti Ghadiyaram and Alan C Bovik. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing* 25, 1 (2015), 372–387.

[17] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1220–1230.

[18] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.

[19] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know. *Trans. Assoc. Comput. Linguistics* 8 (2020), 423–438. doi:10.1162/TACL_A_00324

[20] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. 2020. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 633–651.

[21] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

[22] Lukáš Krasula, Yoann Baveye, and Patrick Le Callet. 2019. Training objective image and video quality estimators using multiple databases. *IEEE Transactions on Multimedia* 22, 4 (2019), 961–969.

[23] Daekyu Kwon, Dongyoung Kim, Sehwan Ki, Younghyun Jo, Hyong-Euk Lee, and Seon Joo Kim. 2024. ATTIQA: Generalizable Image Quality Feature Extractor using Attribute-aware Pretraining. In *Proceedings of the Asian Conference on Computer Vision*. 4526–4543.

[24] Eric C Larson and Damon M Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* 19, 1 (2010), 011006–011006.

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 3045–3059.

[26] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[27] Dingquan Li, Tingting Jiang, and Ming Jiang. 2021. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision* 129, 4 (2021), 1238–1257.

[28] Xiang Lisa Li and Percy Liang. [n. d.]. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 4582–4597.

[29] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 17612–17625.

[30] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–3.

[31] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[32] Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1871–1880.

[33] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[34] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[35] Jiliang Ma, Yihua Chen, Lv Chen, and Zhenjun Tang. 2024. Dual-attention pyramid transformer network for No-Reference Image Quality Assessment. *Expert Systems with Applications* (2024), 125008.

[36] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. 2019. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2344–2348.

[37] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2022. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* 31 (2022), 4149–4161.

[38] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.

[39] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language Models as Knowledge Bases?. In *EMNLP-IJCNLP, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 2463–2473.

[40] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication* 30 (2015), 57–77.

[41] Vishnu Prabhakaran and Gokul Swamy. 2023. Image Quality Assessment using Semi-Supervised Representation Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 538–547.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[45] Subhadeep Roy, Shankhanil Mitra, Soma Biswas, and Rajiv Soundararajan. 2023. Test Time Adaptation for Blind Image Quality Assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16742–16751.

[46] Avinab Saha, Sandeep Mishra, and Alan C Bovik. 2023. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition. 5846–5855.

[47] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[48] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[49] Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained Language Models Yield Few-Shot Semantic Parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 7699–7715.

[50] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. 2023. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*. PMLR, 31716–31731.

[51] Ankit Shukla, Avinash Upadhyay, Swati Bhugra, and Manoj Sharma. 2024. Opinion unaware image quality assessment via adversarial convolutional variational autoencoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2153–2163.

[52] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[53] Suhas Srinath, Shankhanil Mitra, Shika Rao, and Rajiv Soundararajan. 2024. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 22–31.

[54] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3667–3676.

[55] Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. 2023. GraphIQA: Learning Distortion Graph Representations for Blind Image Quality Assessment. *IEEE Trans. Multim.* 25 (2023), 2912–2925. doi:10.1109/TMM.2022.3152942

[56] Wei Sun, Huiyu Duan, Xiongkuo Min, Li Chen, and Guangtao Zhai. 2022. Blind quality assessment for in-the-wild images via hierarchical feature fusion strategy. In *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 01–06.

[57] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. 2023. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing* 17, 6 (2023), 1178–1192.

[58] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[59] Louis L Thurstone. 2017. A law of comparative judgment. In *Scaling*. Routledge, 81–92.

[60] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. 2020. Deep learning on image denoising: An overview. *Neural Networks* 131 (2020), 251–275.

[61] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. 2007. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 383–390.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[63] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. CLIPasso: semantically-aware object sketching. *ACM Trans. Graph.* 41, 4 (2022), 86:1–86:11.

[64] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. 2024. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3635–3647.

[65] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, 2555–2563.

[66] Juan Wang, Zewen Chen, Chunfeng Yuan, Bing Li, Wentao Ma, and Weiming Hu. 2023. Hierarchical curriculum learning for no-reference image quality assessment. *International Journal of Computer Vision* 131, 11 (2023), 3074–3093.

[67] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. 2023. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*. Springer, 46–57.

[68] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. 2024. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19551–19562.

[69] Chenxi Yang, Yujia Liu, Dingquan Li, et al. 2024. Exploring Vulnerabilities of No-Reference Image Quality Assessment Models: A Query-Based Black-Box Method. *arXiv preprint arXiv:2401.05217* (2024).

[70] Dan Yang, Veli-Tapani Peltoketo, and Joni-Kristian Kamarainen. 2019. CNN-based cross-dataset no-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[71] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1191–1200.

[72] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3575–3585.

[73] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. [n. d.]. TF-CLIP: Learning Text-Free CLIP for Video-Based Person Re-identification. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI pages = 6764–6772, publisher = AAAI Press, year = 2024*.

[74] Guanghui Yue, Di Cheng, Leida Li, Tianwei Zhou, Hantao Liu, and Tianfu Wang. 2022. Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network. *IEEE Transactions on Multimedia* (2022).

[75] Guangtao Zhai and Xiongkuo Min. 2020. Perceptual image quality assessment: a survey. *Science China Information Sciences* 63 (2020), 1–52.

[76] Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

[77] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. 2022. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 2864–2878.

[78] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2020. Learning to blindly assess image quality in the laboratory and wild. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 111–115.

[79] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing* 30 (2021), 3474–3486.

[80] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2024. Task-specific normalization for continual learning of blind image quality models. *IEEE Transactions on Image Processing* (2024).

[81] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14071–14081.

[82] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2023. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 440–445.

[83] Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. 2023. Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

[84] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. 2023. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22302–22313.

[85] Yan Zhong, Xingyu Wu, Li Zhang, Chenxi Yang, and Tingting Jiang. 2024. CausalIQA: Towards the Generalization of Image Quality Assessment Based on Causal Inference. In *Forty-first International Conference on Machine Learning*.

[86] Yan Zhong, Chenxi Yang, Suyuan Zhao, and Tingting Jiang. [n. d.]. Semi-Supervised Blind Quality Assessment with Confidence-quantifiable Pseudo-label Learning for Authentic Images. In *Forty-second International Conference on Machine Learning*.

[87] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2023. Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270* (2023).

[88] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 16795–16804. doi:10.1109/CVPR52688.2022.01631

[89] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.* 130, 9 (2022), 2337–2348.

Global Declaration: this supplementary material is the appendix to the main text. The section numbering in this supplementary material follows the numbering in the main text, and the citation references are consistent with those in the main text.

5 Related Works

5.1 Blind Image Quality Assessment (BIQA)

BIQA has gained significant attention recently due to the absence of reference images in realistically distorted image datasets [75]. With the development of Deep Learning (DL), there have sprung up various DL-based methods achieving remarkable progress in BIQA [17, 21, 86], such as QPT [84], GraphIQA [55] and so on. Although these studies can improve the ability of quality perception on training datasets, their generalization is limited by the small scale of the training IQA datasets. To tackle this issue, some works try to improve the generalization of BIQA by larger and more advanced networks [35, 51] or unsupervised pre-training strategies [41, 46], such as HyperIQA [54] and Re-IQA [46, 84]. In addition, some methods are proposed to improve the generalization of IQA models by combining BIQA with various learning paradigms adapted to specific scenarios [1, 23, 53], such as Contrastive Learning, Continual Learning [77], Adversarial Learning [69], Curriculum Learning [66], Multi-task Learning [80] and Causal Inference [85]. More recently, a few studies [65, 81] tried to use the pre-trained CLIP to enhance the generalization of BIQA models, while time-consuming prompt engineering and data scarcity issues still constrain these methods, leading to limited performances. Accordingly, we advance VLM-based IQA with multi-modal mixed-data training and prompt learning. *See Supp. Mat. 5 for more literature reviews on* ***Mixed-datasets Training*** *and* ***CLIP Applications***.

5.2 Prompt Learning

Prompt Learning served as a lightweight knowledge probing technique, prevails in large pre-trained language models (PLMs) to elicit domain-specific knowledge from them with relatively low computational cost [39]. Specifically, the task can be commonly formulated in two formats: (1) *cloze prompts* inducing models to fill in the blank of a textual string [9, 39]; and (2) *prefix prompts* inducing models to complete prefix string [25, 28]. After defining the format of prompts, how to design prompt templates to elicit knowledge effectively becomes a challenge. Pioneering attempts to create prompts are manually designing intuitive templates based on human intelligence. For example, [5] created a series of manually designed prefix prompts in zero-shot, one-shot, and few-shot manners to handle a wide range of downstream NLP tasks. However, manually designing prompts is time-consuming and label-intensive [49], and may fall into sub-optimal [19], constituting a notable limitation in existing several CLIP-based BIQA methods [81]. Therefore, some works aimed to automate the template design process, where continuous prompts are widely adopted in PLMs [6, 28, 83], such as Prefix Tuning [28]. Inspired by these advances, we investigate the performance gains afforded by continuous prompts within the BIQA problems and enhance generalization by our adaptive prompt learning module.

5.3 Mixed Datasets Training

Mixed datasets training favors two main advantages: (1) *one for all*, which provides one unified model for all datasets with better generalization; (2) *all in one*, which utilizes all datasets to train a unified model with better performance. Therefore, many works bring the superiority of mixed datasets training into model learning, such as object-based image classification [8] and semantic segmentation [68]. Due to the high cost and small scale of IQA datasets, it's natural to combine multiple data sets for training, which can significantly improve the generalization of BIQA model. However, one main challenge exists in the inconsistency of subjective quality score ranges across different datasets [22, 27], leading to performance degradation. Consequently, some works try to design effective mixed-dataset training methods for joint IQA training. For example, a Siamese architecture with a cross-dataset training procedure is proposed using the ranking and Euclidean loss [70]. Beyond two datasets, [78] first combines image pairs from individual databases and then trains a BIQA model with image pairs collected from multiple databases simultaneously. Furthermore, to retain the quality space of each database, [57] trains multiple regression modules for each target dataset. However, the pairwise loss will significantly increase the training time, and most of these methods overlook the linear differences between predicted quality scores and ground truth scores. Accordingly, under the premise of multi-modal training, we propose a joint training strategy that incorporates three losses: a relative quality loss, an absolute quality loss, and a linearity-induced loss, with adaptively adjusted weights for each. Our experiments demonstrate that this strategy can greatly enhance the model's generalization ability. It is worth mentioning that, although there have already existed several mixed-data training methods [27, 56], they are all designed for single-modal BIQA approaches. To our knowledge, the training method proposed in this paper is the first multi-modal and adaptive-weight mixed-dataset training method tailored for VLM-based BIQA.

5.4 CLIP Applications

CLIP [43] leveraged large-scale image-text pairs as contrastive views to perform embedding contrast, where the learned visual and textual embeddings have shown high generality and usability in a wide range of visual tasks [50, 63, 64, 73, 88, 89]. Early applications [89] adapt prompt tuning to adapt CLIP to image recognition tasks at a very small cost, and [88] further propose conditional prompt learning to learn generalizable prompts for better out-of-domain performance. Soon after, researchers found CLIP's effectiveness and applicability in a diverse range of tasks, such as object sketching [63], image classification [50], semantic segmentation [64], and video-based person re-identification [73]. Two recent BIQA works [65, 81] closely related to ours intuitively employ the visual and textual embeddings derived from CLIP's encoders to enhance the generalization of BIQA model. Despite effectiveness, these methods suffer from three major issues: (1) neglecting valuable cross-modal features; (2) requiring arduous labor to design prompts; and (3) the small scale of IQA datasets. To cope with them, our work innovatively incorporates multi-modal features to optimize robustness, and we propose a novel adaptive prompt learning method and a comprehensive mixed dataset training strategy in this paper, which can

Table 7: Detailed attributes of IQA databases in experiments.

Databases	Number	Label Range	Distortion Type
TID2013 [40]	3,000	[0, 9]	Synthetic
KADID-10k [30]	10,125	[1, 5]	Synthetic
LIVE [48]	779	[0, 100]	Synthetic
CSIQ [24]	866	[0, 1]	Synthetic
KonIQ-10k [18]	10,073	[1, 5]	Authentic
CLIVE [16]	1,162	[0, 100]	Authentic
SPAQ [14]	11,125	[0, 100]	Authentic
BID [7]	586	[0, 5]	Authentic
PIPAL [20]	29,000	(916, 1836)	AIGC
AGIQA-1k [82]	1,080	[0, 5]	AIGC
AGIQA-3k [26]	2,982	[0, 4.5)	AIGC
AIGCIQA2023 [67]	2400	(26, 73)	AIGC

effectively enhance the generalization performances of the proposed quality-aware CLIP-based BIQA model.

6 More Details Related to Experiments

6.1 Details of Datasets

In this paper, we perform experiments on the following twelve representative IQA databases, including four authentic datasets, four synthetic datasets, and four AIGC datasets, details of which are shown in Table 7:

- TID2013 [40]. This database contains 3000 images, which are obtained from 25 reference images, 24 types of distortions for each reference image, and 5 levels for each type of distortion.
- KADID-10k [30]. It includes 81 pristine images, where each pristine image was degraded by 25 distortions in 5 levels. For each distorted image, 30 reliable degradation category ratings were obtained by crowdsourcing performed by 2,209 crowd workers.
- KonIQ-10k [18]. It consists of 10,073 authentically distorted images chosen from YFCC100M [58] with a wide distribution of brightness, color, contrast, and sharpness.
- CLIVE [16]. CLIVE means the LIVE Challenge Database, consisting of 1162 authentically distorted images captured from many diverse mobile devices. Each image was assessed by an average of 175 unique subjects.
- SPAQ [14]. SPAQ includes 11,125 images taken by 66 mobile phones, which contain a wide range of distortions during shooting, such as sensor noise, blurring due to out-of-focus, motion blurring, over- or under-exposure, color shift, and contrast reduction. And the MOS labels range in $[0, 100]$.
- LIVE [48]. LIVE contains 779 distorted images obtained from 29 pristine images using five different distortion types and more than 25,000 human image quality evaluations. And the DMOS labels range in $[0, 100]$.
- CSIQ [24]. CSIQ contains 886 distorted images created from 30 original images. Six types of distortions are considered in the CSIQ database, which are Gaussian blur, additive white Gaussian noise, JPEG compression, JPEG2000 compression, global contrast decrements, and additive pink Gaussian noise at four or five distortion levels respectively. The DMOS labels range in $[0, 1]$.
- BID [7]. BID is a blur image database that contains 586 images with realistic blur distortion such as out-of-focus, simple motion, complex motion blur, etc. The images were shot under various lighting conditions and exposure times. The labels range in $[0, 5]$.
- PIPAL [20]. PIPAL is an image database from the NTIRE Perceptual Image Quality Assessment Challenge. It includes the results of some image restoration algorithms based on GANs. PIPAL contains 250 high-quality reference images; 40 distortion types, including the output of GAN-based algorithms; 29,000 distortion images and 1,130,000 human ratings. The Elo rating system [13] is employed to assign MOS labels.
- AGIQA-1k [82]. AGIQA-1k is the first perceptual AI-Generated Image (AGI) quality assessment database, which provides 1,080 AGIs along with quality labels, including technical issues, AI artifacts, unnaturalness, discrepancy, and aesthetics as major evaluation aspects.
- AGIQA-3k [26]. AGIQA-3k is a fine-grained AI-Generated Image (AGI) subjective quality assessment database, which consists of 2,982 AGIs generated from 6 different models. This is the first database that covers AGIs from GAN/AR/diffusion-based model altogether.
- AIGC2023 [67]. AIGC2023 includes 2400 AI-Generated Images (AGIs) generated based on 100 various text prompts, and corresponding subjective MOSs evaluated from three perspectives (i.e., quality, authenticity, and text-image correspondence).

For training and testing uniformity, all the label ranges of the four authentic and four synthetic datasets are normalized to $[1, 5]$ during training and testing. The four AIGC datasets are directly tested without label normalization.

6.2 Detailed Experimental Analysis for the Opinion-unaware Methods and PaQ2PiQ

In Section 3.2.1 in the main text, we observe Table 1 and state that Opinion-unaware methods and PaQ2PiQ demonstrate suboptimal performances, and we provide the detailed analysis as follows:

- Opinion-unaware Methods (NIQE, ILNIQE and Ma19): Despite their ambitious goal of handling arbitrary distortions, these methods only perform well on two legacy datasets, LIVE and CSIQ, which feature the basic distortions common in IQA.
- PaQ2PiQ's Limitations: Although trained on a relatively large-scale dataset, PaQ2PiQ shows reasonable quality prediction accuracy mainly on datasets with realistic distortions, but performs poorly on synthetic datasets.

6.3 The Role of Token "Quality" in Template

The token "quality" in the template plays a crucial role in guiding the model to focus on the perceptual quality of images. By explicitly incorporating this token, the model can better align textual descriptions with visual features, enhancing its ability to capture subtle

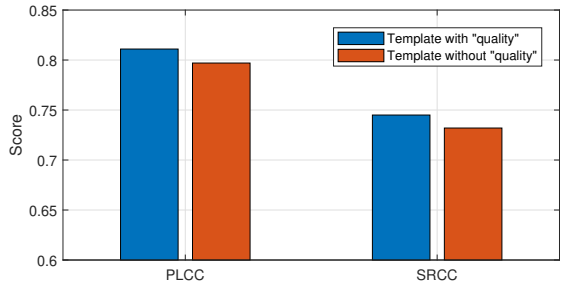


Figure 6: The impact of the 'quality' token in the textual prompts on the proposed MMP-IQA.

quality differences. This is particularly important in multi-modal fusion, where the textual prompt serves as a bridge between visual and linguistic representations, enabling the model to generalize across diverse datasets and distortion types. The inclusion of the "quality" token ensures that the model prioritizes quality-related features during training, leading to more accurate and robust predictions.

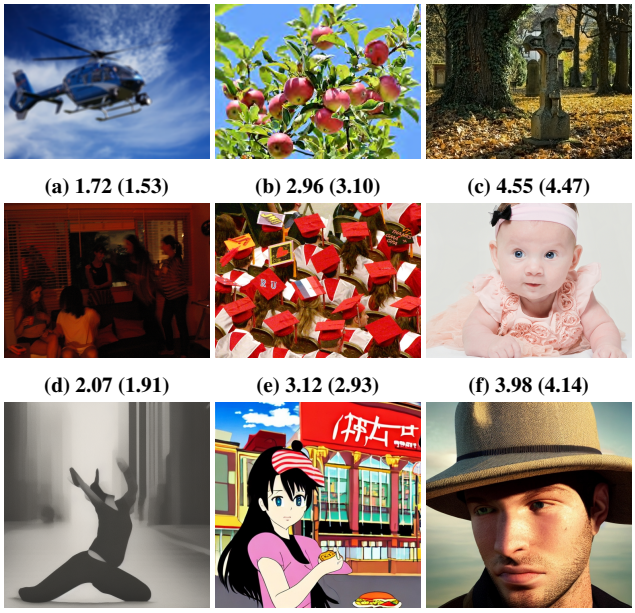
In addition, the another reason we retain "quality" in input text is to form meaningful quality level phrases (e.g., "bad quality", "good quality") that enhance model interpretability for IQA tasks. To validate this design choice, we conducted experiments to compare two templates under identical settings as Figure 2: (1) template with "quality": "A photo with [level] quality"; (2) template without "quality": "A photo with [level]". The results are shown in Figure 6, from which we can observe that retaining "quality" yields better performance compared to its removal, verifying the efficacy of the token "quality" in template during training MMP-IQA.

6.4 The Impact of Different Mixed-Dataset Training Methods on MMP-IQA

Different mixed-dataset training (MDT) methods significantly influence the performance of MMP-IQA. The proposed strategy, which combines relative quality assessment, linearity-induced evaluation, and dataset-specific absolute quality evaluation, allows the model to leverage the strengths of multiple datasets while mitigating their inconsistencies. This approach not only improves generalization but also ensures that the model retains the unique characteristics of each dataset. By adaptively balancing loss weights across datasets and within individual losses, MMP-IQA achieves superior performance compared to traditional single-dataset or fixed-weight training methods. Although the mixed-dataset training method (MDT) in MMP-IQA is not the core contribution, its flexibility enables MMP-IQA to handle complex cross-dataset scenarios, making it more robust and versatile in real-world applications. To validate the effectiveness of our method, we compare it with various mixed-dataset training (MDT) methods, including these settings: (1) MDT_1: directly training by scaling all the MOS ranges to [1,5]; (2) MDT_2: same to the MDT method in [57]; (3) MDT_3: same to the MDT method in [78]; (4) MDT_4: same to the MDT method in [27]. The results are shown in Table-8, proving the superiority of our proposed method.

Table 8: Results of MMP-IQA trained with various MDTs, following the same experimental setups as Table 1 in the main text.

MDTs	MDT_1	MDT_2	MDT_3	MDT_4	Ours
PLCC	0.911	0.909	0.905	0.916	**0.921**
SRCC	0.915	0.924	0.917	0.928	**0.931**



(a) 1.72 (1.53) **(b) 2.96 (3.10)** **(c) 4.55 (4.47)**

(d) 2.07 (1.91) **(e) 3.12 (2.93)** **(f) 3.98 (4.14)**

(g) 1.85 (1.47) **(h) 2.77 (3.10)** **(i) 4.26 (4.61)**

Figure 7: Ranking some examples using our MMP-IQA from the synthetic dataset KADID-10k (a-c), the authentic dataset KonIQ-10k (d-f), and the AIGC dataset AGIQA-1k (g-i). Predicted (and ground truth) scores are shown below each image.

6.5 Some Examples Predicted by MMP-IQA

In Figure 7, we show some examples predicted by our MMP-IQA, from which we can see that the predicted scores are close to ground truth scores. Note that the results of (g-i) on AGIQA-1k are zero-shot results without fine-tuning.

7 More Details Related to Methods

7.1 Details of Cross-modal Fusion in Eq. (7)

Inspired by [87], we use the self-attention (SA) mechanism [31, 62] as the cross-modal fusion function $\mathcal{R}(\cdot)$. Being permutation invariant, SA is good at outputting adapted embeddings even with long dependencies, which naturally suits the characteristics of the adaptation function. Specifically, SA keeps the triplets (query $\mathcal{Q}$, key $\mathcal{K}$, value $\mathcal{V}$) with $\mathcal{Q} = \mathcal{K} = \mathcal{V} = [\mathcal{S}(z), \textbf{Context}]$, where $\textbf{Context} = [\mathcal{W}, \mathcal{P}]$ is the textual context with the visual information ($\mathcal{P} = [\mathcal{P}_1, \ldots, \mathcal{P}_5] \in \mathbb{R}^{5\times K}$ and $\mathcal{W} = [\mathcal{W}_1, \ldots, \mathcal{W}_5] \in \mathbb{R}^{5\times K}$), and $\mathcal{S}(z)$ is the visual representation produced by Meta-IQA. The inputs are projected into the same space, that is:

$$K = W_K^\top [\mathbf{k}_k; \forall \mathbf{k}_k \in \mathcal{K}] \in \mathbb{R}^{d\times|\mathcal{K}|}. \tag{22}$$

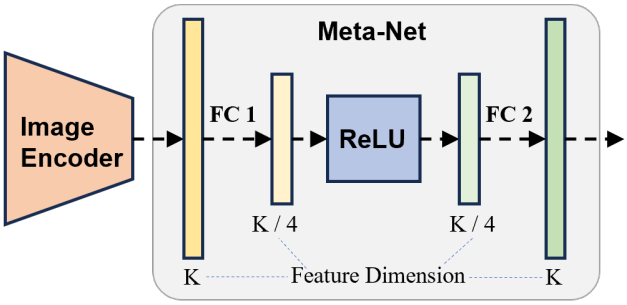


Figure 8: Detailed structure of Meta-Net in MMP-IQA.

Similar projections are made for $\mathcal{Q}$ and $\mathcal{V}$. The query $\boldsymbol{x}_q \in \mathcal{Q}$ is matched against a list of keys K where each key has a value V. The output is the sum of all the values weighted by the proximity of the key to the query point:

$$\tilde{z} = \sum_k \alpha_{qk} V_{:,k}, \quad \alpha_{qk} \propto \exp\left(\frac{\mathcal{S}(z)^\top W_Q \cdot K}{\sqrt{d}}\right) \tag{23}$$

$$\tilde{\textbf{Context}} = \sum_k \beta_{qk} V_{:,k}, \quad \beta_{qk} \propto \exp\left(\frac{\textbf{Context}^\top W_Q \cdot K}{\sqrt{d}}\right) \tag{24}$$

where $V_{:,k}$ is the k-th column of V. $\tilde{\textbf{Context}} = [\tilde{\mathcal{W}}, \tilde{\mathcal{P}}]$ is instance-conditioned context after cross-modal fusion,

7.2 Details of Meta-Net in MMP-IQA

As stated in Section 2.3.2, the Meta-Net includes two FC layers, maintaining consistent input and output dimensions. Specifically, the input image $\boldsymbol{x}$ is first encoded by the image encoder $\boldsymbol{h}_\phi : \mathbb{R}^N \mapsto \mathbb{R}^K$ (with $K = 512$), generating a feature representation $\boldsymbol{z} = \boldsymbol{h}_\phi(\boldsymbol{x})$ that serves as the input to the Meta-Net. As shown in Figure 8, the visual representation $\boldsymbol{z}$ first passes through the first fully connected layer, **FC 1**, which maps the dimension of $\boldsymbol{z}$ from K to $K/4$. After applying the Rectified Linear Unit (ReLU) activation, it then passes through a second fully connected layer, **FC 2**, which restores the dimension of $\boldsymbol{z}$ from $K/4$ back to K.

As shown in Figure 3, although the feature representation directly generated by the image encoder $\boldsymbol{h}_\phi : \mathbb{R}^N \mapsto \mathbb{R}^K$ and the subsequent visual representation produced by Meta-IQA are both extracted during training, they serve different roles and functions. **For the former**, its role is to calculate cosine similarity with the feature representations of five different quality-level text descriptions, thereby deriving the probability of the input image belonging to each quality level. **For the latter**, according to Eq. (7) and Figure 1 (B), its role is to further extract quality-related latent features from the input image and, through the cross-modal fusion module, refine the context that represents the five quality levels into an instance-conditioned context, thereby enhancing the model's generalization capability.

7.3 Comparison between LIQE and MMP-IQA

Since the primary comparison of the proposed MMP-IQA in this paper is with LIQE (which, to our knowledge, is the state-of-the-art CLIP-based BIQA model), we not only compare the experimental results but also analyze the algorithms, focusing on the following key points.

- Given the fact that the image quality is also related to scene content and distortion type, LIQE is a CLIP-based method designed from a multi-task training perspective, while MMP-IQA is another CLIP-based method designed from the information fusion perspective. Unlike LIQE, which requires manual construction of labels for scene content and distortion type, MMP-IQA leverages a MetaNet to mine latent quality-related elements automatically. This enables a more effective instance-conditioned multi-modal fusion module, enhancing the model's generalization ability and avoiding LIQE's limitations in training data labeling.
- LIQE's textual templates are manually designed, which is time-consuming and labor-intensive. In contrast, MMP-IQA introduces an adaptive prompt learning approach that not only can automatically optimize image quality text prompts but also enhance the model's effectiveness.
- LIQE only uses text information to differentiate between quality levels. However, recent works have identified an inherent domain gap [29] between visual and textual embeddings in VLM, causing them to reside in separate clusters in the embedding space and hindering effective pair-wise mapping. Unlike LIQE, MMP-IQA further extracts visual information during training to distinguish between quality levels and fuses it with textual information. This fusion creates new cross-modal textual features for distinguishing between quality levels, which can effectively enhance the model's generalization capabilities.
- To train on multiple datasets, LIQE utilizes pairwise loss to address the inconsistency of subjective quality score ranges across different datasets. However, in addition to aligning subjective quality scores, it is also crucial to maintain the unique characteristics of each dataset. Therefore, MMP-IQA not only employs pairwise loss but also incorporates a dataset-specific absolute quality evaluator to retain the distinct quality space of each dataset, effectively enhancing the performance of mixed dataset training. Moreover, in LIQE, each dataset's loss is treated equally during training, which limits the performance when training on mixed datasets. To address this issue, we propose a multi-modal adaptive weighting method for mixed dataset training in MMP-IQA, effectively enhancing the model's generalization capability.

Note that in Section 3.2.1 in the main text, we compared the parameter counts (LIQE: 151.3*M* **vs.** MMP-IQA: 152.5*M*) and training speeds (LIQE: 42.6 *samples / sec* **vs.** MMP-IQA: 40.2 *samples / sec*) on a single NVIDIA A10 GPU. The results demonstrate comparable computational scales between the two models. We emphasize that identical experimental configurations were adopted during testing: the *num worker* was set to 8, and batch sizes were consistently set to 16 for KADID-10k and KonIQ-10k, and to 4 for LIVE, CSIQ, BID, and CLIVE datasets. The number of training steps in each epoch was both set as 200 in MMP-IQA and LIQE.

8 Proof of Theorem 2.1 and Correlative Analysis

First, we have the following definitions:

Definition 8.1 (Generalization Error [2]). In statistical learning and machine learning, the generalization error refers to the difference between a model's performance on training data (training error) and its performance on unseen test data (test error). Formally, for a loss function $L_k^{(d)}$:

$$\epsilon_k^{(d)} = \left| L_k^{(d)} - \mathbb{E}[L_k^{(d)}] \right|, \tag{25}$$

where $L_k^{(d)}$ is the empirical loss on training data $\mathbb{E}[L_k^{(d)}]$ is the expected loss over the true data distribution

Definition 8.2 (Hoeffding's Inequality [3]). Let $Z_1, \ldots, Z_{N_d}$ be independent random variables bounded by $a \le Z_i \le b$. For any $\epsilon > 0$, the following inequality holds:

$$\mathbb{P}\left(\left| \frac{1}{N_d} \sum_{i=1}^{N_d} Z_i - \mathbb{E}[Z] \right| \ge \epsilon \right) \le 2 \exp\left(-\frac{2N_d \epsilon^2}{(b-a)^2} \right). \tag{26}$$

Hoeffding's inequality provides concentration bounds for the sum of independent bounded random variables.

Then we can prove Theorem 2.1 as follows:

PROOF. According to Eq. (18) and Eq. (19), the joint loss function of MMP-IQA is:

$$L_{\text{total}} = \sum_{k=1}^{3} \lambda_k \sum_{d=1}^{D} w_k^{(d)} L_k^{(d)}, \tag{27}$$

where D is the number of datasets $\mathcal{D}_1, \ldots, \mathcal{D}_D$, λ_k are outer weights (between different loss terms), $w_k^{(d)}$ are inner weights (between datasets for the same loss term), and $L_k^{(d)}$ is the k-th loss function on dataset $\mathcal{D}_d$ ($k = 1, 2, 3$ for relative, linearity-induced, and absolute quality assessment respectively)

Then, the generalization error of joint loss decomposes as:

$$\mathbb{E}[L_{\text{total}}] = \sum_{k=1}^{3} \lambda_k \sum_{d=1}^{D} w_k^{(d)} \mathbb{E}[L_k^{(d)}] \tag{28}$$

In our proposed MMP-IQA method, the empirical loss $L_k^{(d)} = \frac{1}{N_d} \sum_{i=1}^{N_d} \ell_k(x_i, y_i)$, where ℓ_k is the per-sample loss. Since all the label ranges of the authentic and synthetic datasets used in our experiments are normalized (as stated in Supp. Mat. 6.1), ℓ_k is bounded: $0 \le \ell_k \le M$ (i.e., in MMP-IQA, quality scores are typically normalized to $[1, 5]$). Note that the expected loss $E[L_k^{(d)}] = E_{x,y}[\ell_k(x, y)]$. Assume each loss function $L_k^{(d)}$ is M-Lipschitz continuous and bounded (valid for BIQA as quality scores are bounded), let $a = 0, b = M$, according to the Hoeffding's inequality in Definition 8.2, we have:

$$\mathbb{P}\left(\left| L_k^{(d)} - \mathbb{E}[L_k^{(d)}] \right| \ge \epsilon \right) \le 2 \exp\left(-\frac{2N_d \epsilon^2}{M^2} \right). \tag{29}$$

To find an ϵ such that the probability on the right-hand side of the inequality in Eq. (29) is sufficiently small (e.g., $\le \delta$, $\delta = 0.05$ for 95% confidence), that is:

$$2 \exp\left(-\frac{2N_d \epsilon^2}{M^2} \right) \le \delta. \tag{30}$$

Solving the inequality in Eq. (30), we have:

$$\exp\left(-\frac{2N_d \epsilon^2}{M^2} \right) \le \frac{\delta}{2} \implies -\frac{2N_d \epsilon^2}{M^2} \le \ln\left(\frac{\delta}{2} \right). \tag{31}$$

Further solving yields:

$$\epsilon \ge M \sqrt{\frac{\ln(2/\delta)}{2N_d}}. \tag{32}$$

Let $C_k = M\sqrt{\frac{\ln(2/\delta)}{2}}$, then $\epsilon = \frac{C_k}{\sqrt{N_d}}$. Therefore, with probability at least $1 - \delta$, the following inequality holds:

$$\left| L_k^{(d)} - \mathbb{E}[L_k^{(d)}] \right| \le \frac{C_k}{\sqrt{N_d}}. \tag{33}$$

After the substitution from Eq. (33) to Eq. (28), we obtain:

$$\mathbb{E}[L_{\text{total}}] \le \sum_{k=1}^{3} \lambda_k \sum_{d=1}^{D} w_k^{(d)} \left(L_k^{(d)} + \frac{C_k}{\sqrt{N_d}} \right). \tag{34}$$

Note that the inner weights $w_k^{(d)}$ are normalized via softmax, and the outer weights λ_k are adjusted according to relative descent rates:

$$w_k^{(d)} = \frac{e^{L_k^{(d)}}}{\sum_{d'=1}^{D} e^{L_k^{(d')}}}, \quad \lambda_k(t) = \frac{\exp(\beta_k(t-1)/\tau_3)}{\sum_{j=1}^{3} \exp(\beta_j(t-1)/\tau_3)}. \tag{35}$$

Then we have:

$$\sum_{d=1}^{D} w_k^{(d)} = 1, \quad \sum_{k=1}^{3} \lambda_k = 1. \tag{36}$$

Thus, when training stabilizes, both λ_k and $w_k^{(d)}$ satisfy the normalization conditions. Consequently, we can derive the upper bound for the generalization error: Firstly, let $C = \max_k C_k$ and utilize the weight normalization properties, we have:

$$\mathbb{E}[L_{\text{total}}] \le \sum_{k=1}^{3} \lambda_k \sum_{d=1}^{D} w_k^{(d)} L_k^{(d)} + \sum_{k=1}^{3} \lambda_k \sum_{d=1}^{D} w_k^{(d)} \frac{C}{\sqrt{N_d}}. \tag{37}$$

Given that $\sum_{k=1}^{3} \lambda_k = 1$ and $\sum_{d=1}^{D} w_k^{(d)} = 1$, the first term represents the empirical loss L_{total}, while the second term constitutes the generalization error component:

$$\mathbb{E}[L_{\text{total}}] \le L_{\text{total}} + \sum_{d=1}^{D} \frac{C}{\sqrt{N_d}} \left(\sum_{k=1}^{3} \lambda_k w_k^{(d)} \right). \tag{38}$$

By disregarding the empirical loss L_{total} (since the optimization objective is to minimize it), according to Definition 8.1, we obtain the upper bound for generalization error G_e:

$$G_e = \mathbb{E}[L_{\text{total}}] - L_{\text{total}} \le O\left(\sum_{d=1}^{D} \frac{1}{\sqrt{N_d}} \left(\sum_{k=1}^{3} \lambda_k w_k^{(d)} \right) \right). \tag{39}$$

Q.E.D. □

From the theoretical results in Eq. (39), we provide the theoretical justification for mixed dataset training stability:

- **Dominance of Sample Size N_d.** Following from the Law of Large Numbers and Hoeffding's inequality for bounded loss functions, as N_d increases, $\epsilon_k^{(d)}$ decreases at rate $1/\sqrt{N_d}$. The generalization error bound is inversely proportional to $\frac{1}{\sqrt{N_d}}$, which implies: (1) Larger datasets (with bigger N_d) contribute

less to the generalization error because $\frac{1}{\sqrt{N_d}}$ decays as N_d increases. This ensures that models prioritize learning from statistically reliable data sources; (2) smaller datasets (with limited N_d) significantly increase the upper bound of generalization error due to larger $\frac{1}{\sqrt{N_d}}$ values. Consequently, the overall generalization error is primarily governed by the smallest dataset in the training mixture. Therefore, the adaptive weighting system dynamically balances contributions across datasets and tasks without manual intervention. This implicit normalization handles heterogeneous data distributions and prevents any single dataset or task from monopolizing the training process.

- **Regulatory Role of Adaptive Weighting.** Inner weights $w_k^{(d)}$ automatically balance contributions across datasets: Higher weights for datasets with larger losses, prioritizing their optimization; lower weights for well-performing datasets to avoid excessive tuning. Outer weights λ_k balance contributions from different loss terms (e.g., relative evaluation vs. absolute evaluation): If a loss term decreases rapidly, its weight will be reduced to avoid dominating the training process; terms with slower loss reduction will receive higher weights to ensure multi-task balance. Therefore, the adaptive weighting system dynamically balances contributions across datasets and tasks without manual intervention. This implicit normalization handles heterogeneous data distributions and prevents any single dataset or task from monopolizing the training process.
- **Implicitly Normalizing Data Distribution.** Since different datasets may use varying MOS ranges (e.g., [0,100] vs. [1,5]), causing model confusion if directly combined. To address this issue, in MMP-IQA, relative quality assessment (L_1) learns ranking relationships across datasets, ignoring absolute value differences; linearity-induced loss (L_2) enforces linear consistency between predicted and subjective scores; and dataset-specific absolute quality assessment (L_3) preserves unique characteristics of each dataset. Therefore, the combined effect of adaptive weighting and multi-loss design implicitly aligns label ranges into a unified feature space, eliminating the need for explicit normalization or score scaling while maintaining dataset-specific traits.