MMIFN: A Multi-Modal Interactive Fusion Network for Omnidirectional Image Quality Assessment

Yun Liu, *Member, IEEE,* Sifan Li, Huiyu Duan, Daoxin Fan,and Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Omnidirectional image quality assessment (OIQA) has attracted increasing attention in recent years but is still challenging due to its complexity. Most earlier works are built based on visual degradation-aware features from the entire distorted image or/and the viewports, which focus on extracting representative visual features. Unlike conventional OIQA methods only relying on visual information, we propose an effective multimodal interactive fusion network for OIQA with the help of textual semantic information, which is inspired by the non-declarative and declarative memory cognition mechanism in human brain. Specifically, the text information is firstly captured based on the available visual modality to get a detailed and accurate quality description, which can provide rich declarative semantic information, and complement with the non-declarative visual information. Then the textual semantic information is not only fused with the global visual features at various levels but also fused with the local visual features following human view characteristics of omnidirectional image. Finally, an aggregated integrated module is designed to further strengthen the feature fusion, which digs the deep image-text interaction relation and improve the text-image fusion accuracy. Extensive experiments on two public datasets prove the effectiveness of our model, and the benefit of multimodal information on OIQA field.**

***Index Terms*—OIQA, image quality assessment, image quality, multimodal, neural network.**

I. Introduction

NOWADAYS, omnidirectional image (OI) plays a significant role in modern social media, film, and entertainment, and so on [1], [2]. Similar to two-dimensional (2D) image, the quality of the omnidirectional image is influenced by several factors, including capturing method, compression model, and so on. Different from traditional 2D image, the omnidirectional image has a large field of view (FoV) and can provide immersive visual representation of the environment with some auxiliary equipments, e.g., Head Mounted Display (HMD). OI usually captures as the spherical formats, and need to projects to 2D plane for transmission and storage, such as equirectangular projection (ERP) format and cube-map projection (CMP) format. There is significant difference between OI and traditional 2D image, which makes that 2D image quality assessment method fail to predict the quality of omnidirectional image [3]–[7]. The 360° × 80° FoV and special projection format makes the omnidirectional image quality assessment (OIQA) challenge [8].

Generally, OIQA is divided into subjective method and objective methods. Subjects have to ware HMD to watch an OI, and each time they only watch one viewport, which takes much time for one OI [9]. To avoid fatigue during subjective experiments, subjects have to take a long time rest. Since subjective method is time-consuming, labor-intensive, and costly, objective method arouses much attention. To effectively evaluate the quality of the omnidirectional image, researchers designed three types of objective methods, namely full-reference (FR), reduced-reference (RR) and no-reference (NR) metrics [14]–[16]. FR and RR methods, they all need the information of the reference image. However, in real application area, the reference image is not always available, so NR models became the mainstream methods [10]–[13]. At the early stage, NR omnidirectional image quality assessment (OIQA) methods mainly designed based on handcraft features, such as color, structure, texture and other attributes, which are extracted from the entire OI [15], [17], [18]. The above works, on one hand, took the 2D feature extraction methods on OI and neglect the characteristics of OI, and on the other hand, the performance is highly relied on the extracted handcraft features, which limits the overall accuracy. Then some researchers focus on the experience procedure of OI, and notice that what users watching is the contents in each viewport, which provides a new direction of extracting important visual features. The patch-based model and viewport-based model are proposed to provide rich local visual features. Considering the complementation mechanism between the local and global information, many studies are proposed based on combing the local and global visual features to improve the performance of OIQA model [18]–[20]. With the development of deep learning technology, researchers then introduce many effective networks to predict the quality of OI [21]–[23], which presents a promising performance.

With the advancement of deep learning technology, deep

This work was supported in part by Shenyang science and technology plan project under Grant 23-407-3-32, in part by Natural Science Foundation of Liaoning Province under Grant 2023-MS-139, and in part by National Natural Science Foundation of China under Grant 61901205. *(Corresponding author: Yun Liu.)*

Yun Liu, Sifan Li, and Daoxin Fan are with the Faculty of Information, Liaoning University, Shenyang 110036, China (e-mail: yunliu@lnu.edu.cn; sflijohn@foxmail.com; fdx_0729@163.com).

Huiyu Duan and Guangtao Zhai are with the Institute of Image Communication and Information Processing, Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: huiyuduan@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

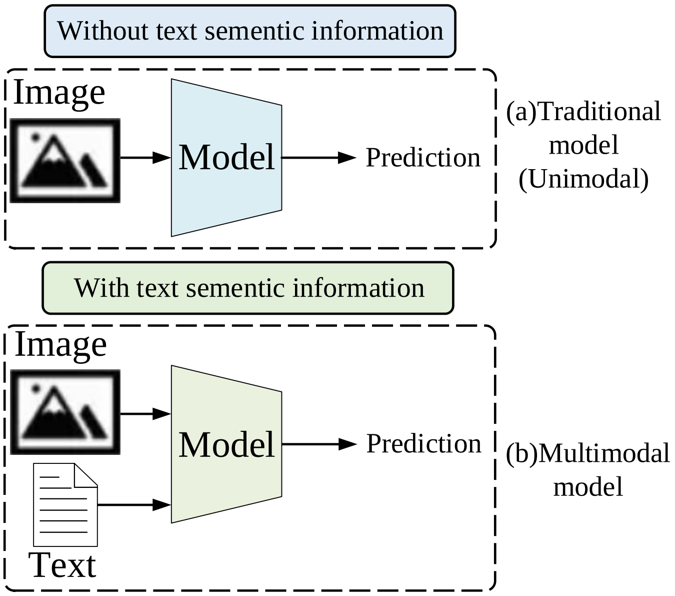

Without text sementic information
Image
Model
Prediction
(a)Traditional model (Unimodal)
With text sementic information
Image
Model
Prediction
Text
(b)Multimodal model

Fig. 1. The difference of traditional image quality assessment metric and multimodal image quality assessment.

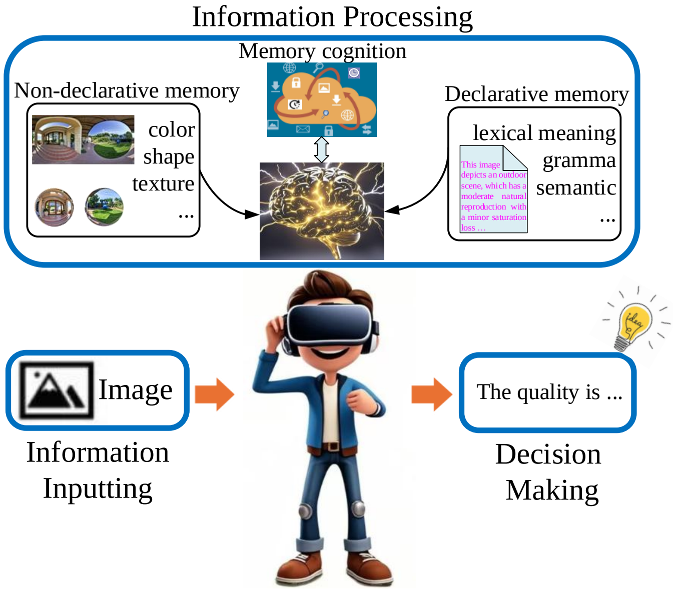


Fig. 2. The diagram of decision-making mechanism, including information inputting, memory cognition and decision making, which indicates the fusion mechanism of declarative text semantics and non-declarative image semantics in human brain.

learning-based models then show their strong ability on various visual tasks and became the mainstream IQA methods, including 2D, 3D and OI fields. On the basis of previous works, Xu *et al.* [24] and Zhang *et al.* [25] proposed some effective networks by simulating human visual system. Although local and global visual features are both considered, multi-scale visual features and interactive relationship between features, which can provide rich detail information, are neglected. Later, Zhang *et al.* [26], based on scene content understanding, proposed a multi-scale feature fusion model based on the saliency visual attention mechanism, which proves our hypothesis of the important of multi-scale visual features on OIQA. Although the above works achieve a good improvement, great efforts are still needed to build more effective models, especially on digging multi-scale visual features and rich semantic information.

With the development of information technology, the modern digital content has quickly evolved to multimodality. For example, images on social media are often companied with text, and people are allowed to share their comments on images or videos. The textual information can provide additional complementary information to help people better understand the contents [27]. Emerging research has explored the potential of text modality in quality assessment, and proposed many multimodal models, which yield good performance on various computer vision tasks [28], [29]. Li *et al.* [30] considered the interactions between visual and textual modalities and presented an attribute-assisted multimodal memory network (AMM-Net) for image aesthetics assessment. Zhang *et al.* [31] designed an effective point cloud quality assessment model through text supervision, which proves the benefit of multimodal information. Considering that the other modality is not always available, Ma *et al.* [32] reconstructed the missing modalities by using the Bayesian meta-learning framework and effectively solved the problem. Zhang *et al.* [33] focused on the problem of missing textual modality in image aesthetic quality assessment, they proposed a novel multimodal image aesthetic quality assessment method by reconstructing the missing textual description according to the available image information. Xie *et al.* [27] then utilized the large language model (LLM) to generate quality description texts for 3D object and proposed a point cloud quality assessment approach based on graph learning, which further proves the importance of text information in IQA area. As shown in Fig. 1, in contrast to traditional image quality assessment model, the above works not only capture the visual features of image, but also the quality-aware semantic information in the text. Semantic information greatly influenced quality experience of humans. Humans usually percept distortions, identify the quality degradations of these distortions, like color, texture distortion, and then conclude the quality-aware semantic information to predict the quality. Traditional models just rely on the visual features, which cannot be effectively modeled the reasoning process in human brain. While the text description can prove rich semantic information and well complement with image visual information, which has been proved by the above works and provides us with a new way to design an effective OIQA model.

Besides, cognitive psychology found that there exists declarative and non-declarative memory cognition in human brain to help people make decisions [36]. Based on neuroscience research, this process can be explained from three key stages: information inputting, memory cognition, and decision making. Take the OI quality assessment as an example, as shown in Fig. 2. Human eyes receive image information and convert it into neural signals by the retina, and then the signals are transmitted to human brain. As the classical non-declarative information, image visual signal is recognized through the primary visual cortex of the occipital lobe, the inferior temporal gyrus, etc. The classical declarative information is further recognized by the visual word form area, and the semantic parsing is completed in combination with the Broca's area, Wernicke's area, etc. Subsequently, declarative text semantics and non-declarative image semantics blend in the semantic memory networks of the temporal lobe and the prefrontal lobe to form the cognition memory. Finally, the brain integrates multi-source information based on the dynamic interaction of multi-brain neural networks and accomplishes the evaluation system and decision-making. Inspired by the above decision-making mechanism, it is reasonable to combine text information and image information to conduct the OI quality prediction. The information hidden in image and text, and the interactive relationship between them can help models learn significant quality-aware information.

In this paper, we propose an effective multimodal interactive fusion network for omnidirectional image quality assessment, namely MMIFN, which not only considers the quality degradation-aware information hidden in each modal, but also the multimodal interactive relationship related to quality degradation. Firstly, the descript text is generated by using a novel multi-modal large language model-based method, DepictQA, which can describe the distortions, texture damages and the overall quality of the distorted image, aligning closely with human's reasoning process [37]. Then the visual perception features and the textual semantic features are both extracted for further processing. Considering the complex connection of cognition memory mechanism in human brain [35], the textual semantic feature is fused with the global visual features and local visual features, respectively, based on two effective cross-modal interaction modules, which can capture the interactive relationship between modalities. Finally, an aggregated integrated module is designed to fuse the multimodal fusion features and output the predicted quality score. The main contributions are as follows:

- The declarative semantic information provided by the text description can well complement the non-declarative visual perceptual information. To the best of our knowledge, it is the first time to introduce text information to the field of omnidirectional image quality assessment.
- A global feature fusion (GFF) module is built to dig the multi-scale global inherent correlation between visual features and textual semantic information, which can capture the representative global quality-aware features to improve the overall performance gain over existing state-of-the-art methods.
- To obtain key semantic details and improve the text-image fusion accuracy, local visual features and textual semantic information are fine-grained fused to complement the global fusion features based on a local feature fusion (LFF) module, which can improve multimodal feature learning ability and the accuracy of our model.
- An aggregated fusion module (AFM) is designed to smoothly fuse the global and local multimodal features, which can achieve robust and accurate prediction.

We arrange the remainder of this paper as follows. The related works are briefly summarized in Section II. Our proposed method is introduced in detail in Section III, and the experimental results and the analysis are reported in Section IV. Finally, we present the conclusion of this work in Section V.

II. Related Works

This section reviewed the related works from the following two aspects: (1) OIQA models and (2) text semantic information.

A. OIQA Models

At the early stage, OIQA models are built based on some successful 2D metrics in 2D IQA area, such as SSIM, PSNR, and so on [38]–[40]. However, by directly applying 2D IQA metrics to the whole OI, the above models cannot effectively predict the quality of OI. To seek for effective quality-aware visual feature extracting method, the patch-based models are proposed based on the projection format of OI, such as segmented spherical projection (SSP) format, the cubemap projection (CMP) format, and equirectangular projection (ERP) format, etc. [41], [42]. To handle the distortion problem in the equatorial region of the ERP format, Zheng *et al.* [35] designed an NR OIQA metric based on the local details and global features. Jiang *et al.* [20] proposed an omnidirectional image quality assessment model by combining global visual features and local visual features extracted from non-overlapping patches, which gets a better performance than the above models. Ma *et al.* [43] proposed a mutual distillation based OIQA framework by exploring the complementary relationship between different projection formats. However, the patch-based methods do not coincide with human vision perception mechanisms while people watch an omnidirectional image.

Considering that people can only see the limited contents of OI at one time, the viewport-based OIQA model are proposed [44]–[46]. Sun *et al.* [47] developed an effective NR-OIQA model by combing the local visual features of viewports, which proves the importance of viewports' features in OIQA task. Xu *et al.* [24] built an omnidirectional image quality assessment with viewport-oriented graph convolutional networks. Motivated by the effectiveness of saliency perception characteristics in 2D IQA task, Zhang *et al.* [26] proposed a saliency-guided NR OIQA method by fusion multi-scale features of each viewport. Overall, the above viewport-based models achieve better performance than the patch-based models, and the viewport-based models combing the local and global features present a competitive result, which proves the reasonable of combining global perception information and the local perception information.

B. Text Semantic Information

Recently, people uploaded textual comments on the internet to share their subjective perceptual information of images, which promotes the emergence of multimodal models in many tasks, such as image aesthetic assessment (IAA) task and point cloud quality assessment task, which presents a

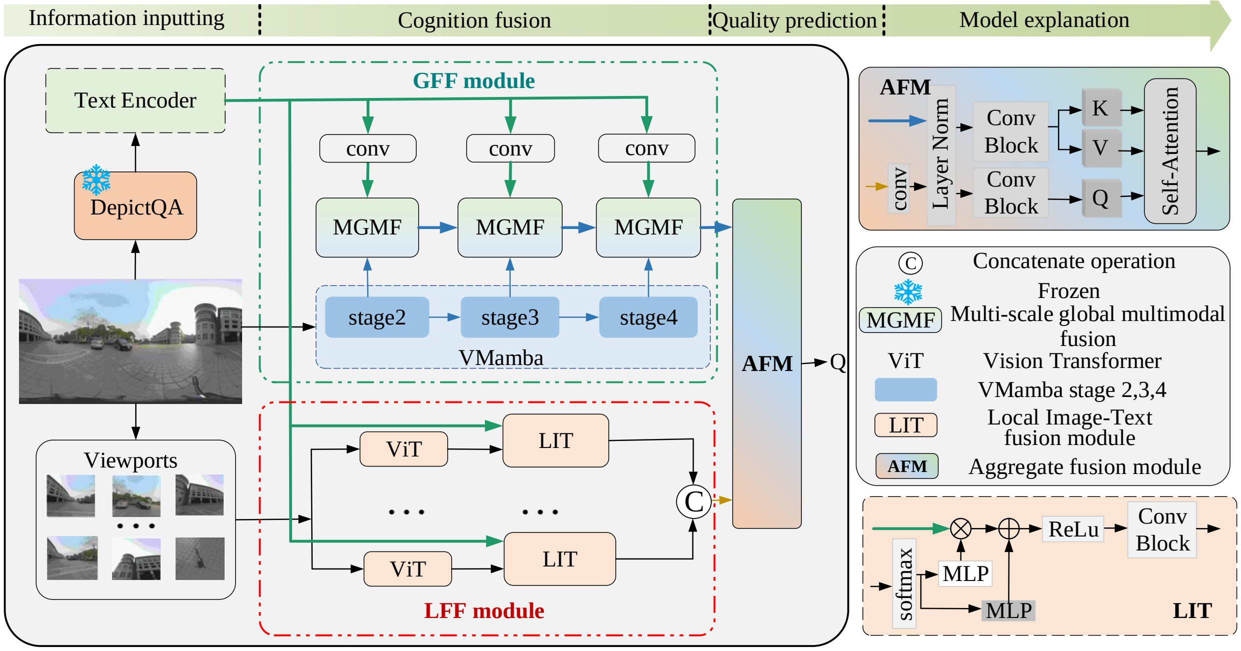


Fig. 3. The flowchart of the proposed model, which includes information inputting, cognition fusion, and quality prediction. We first obtain the viewports of the distorted OI and its text description, and then the text semantic information is fused with the non-declarative global and local visual features for cognition. Finally, an aggregated fusion module based on attention mechanism is used for quality prediction.

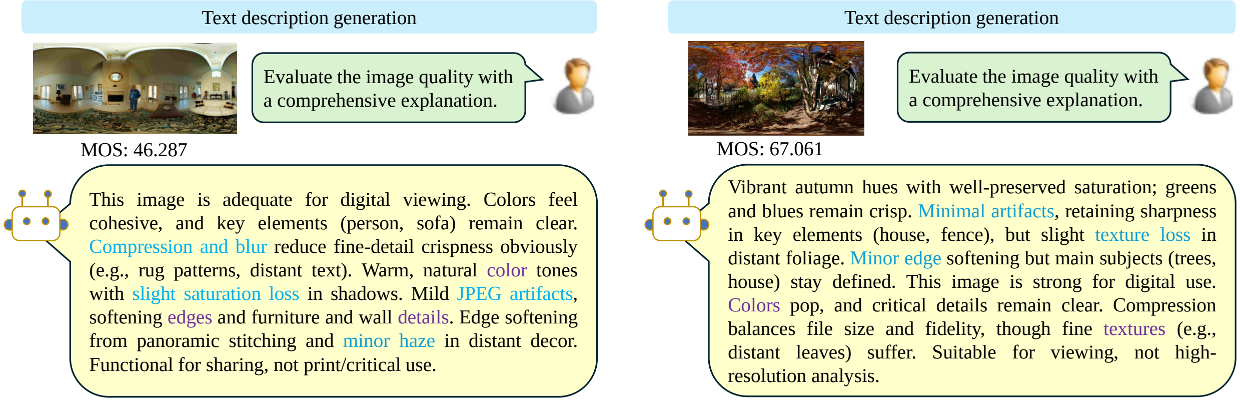
promising result. Compared to the unimodal IAA model, the multimodal IAA models get a better performance with the help of text semantic information [35], [48], which proves the importance of text information in IAA tasks and the necessity of combining the text semantic information. However, most existing image datasets only give the quality score without the text information, which makes the combing problem challenge. Fortunately, some methods have been proposed to solve the text generation problem [39], [49]. Motivated by the above works, Zhang *et al.* [50] utilized visual information to reconstruct missing textual information and proposed an image aesthetic assessment network. Inspired by recent Large Language Models (LLMs) and multi-modal technologies, You *et al.* [37] design a new paradigm for IQA, named Depicted image Quality Assessment (DepictQA), to interpret image content and distortions descriptively and comparatively, which provide us a novel way to generate accurate quality descriptions to build effective OIQA model.

III. Proposed Method

In this paper, we propose a novel multimodal omnidirectional image quality assessment model guided by human brain memory mechanism, which can be concluded into three stages: information inputting, cognition fusion, and quality prediction. The overall framework is present in Fig. 3. For the information inputting stage, we utilize the pretrained DepictQA model [37] to generate the missing text description to obtain the complete multimodal information and the declarative semantic information, and the available distorted OI is segmented into several viewports to capture the local visual information. For the cognition fusion stage, following human cognition memory mechanism and characteristics of OI, the declarative semantic information is not only fused with non-declarative global visual features, but also non-declarative local visual features based on two multimodal fusion models, GFF module and LFF module. Finally, for the quality prediction stage, we design an aggregated fusion module, AFM, to further strengthen the feature fusion and map to the quality score. The details are as follows.

A. Information Inputting

1) Text Generation and Semantic Extraction: Although applying the text information to solve vison task has been developed for a certain time, it has not introduced to OIQA field. Besides, the apparent challenge is that exited OI datasets are provided without text information, which further delays the development of multimodal model on OIQA. In order to deal with the missing text information, we reconstruct the missing textual description related to the visual quality according to the available entire OI. In this paper, we apply the DepictQA [37] model to output the text description to describe the quality of OI with a comprehensive explanation, which can be naturally understood by humans and greatly improving the interpretability [51]. Through sending the distorted OI to DepictQA with the text prompt: *Evaluate the image quality with a comprehensive explanation*, DepictQA can generate a detailed and accurate quality description, as shown in Fig. 4. As we can see, for the given distorted OI, DepictQA gives a detailed quality description from different aspects that humans are usually concerned about. The quality degradation description is related to color, texture, distortion content and so on, and images with different quality are given different text description, which proves that DepictQA can provide distinctive quality description for different OI. The



Evaluate the image quality with a comprehensive explanation.

This image is adequate for digital viewing. Colors feel cohesive, and key elements (person, sofa) remain clear. Compression and blur reduce fine-detail crispness obviously (e.g., rug patterns, distant text). Warm, natural color tones with slight saturation loss in shadows. Mild JPEG artifacts, softening edges and furniture and wall details. Edge softening from panoramic stitching and minor haze in distant decor. Functional for sharing, not print/critical use.

Evaluate the image quality with a comprehensive explanation.

Vibrant autumn hues with well-preserved saturation; greens and blues remain crisp. Minimal artifacts, retaining sharpness in key elements (house, fence), but slight texture loss in distant foliage. Minor edge softening but main subjects (trees, house) stay defined. This image is strong for digital use. Colors pop, and critical details remain clear. Compression balances file size and fidelity, though fine textures (e.g., distant leaves) suffer. Suitable for viewing, not high-resolution analysis.

Fig. 4. The text description generated based on DepictQA [37], which can describe the distortions, texture damage and the overall quality of the distorted image, aligning closely with human's reasoning process.

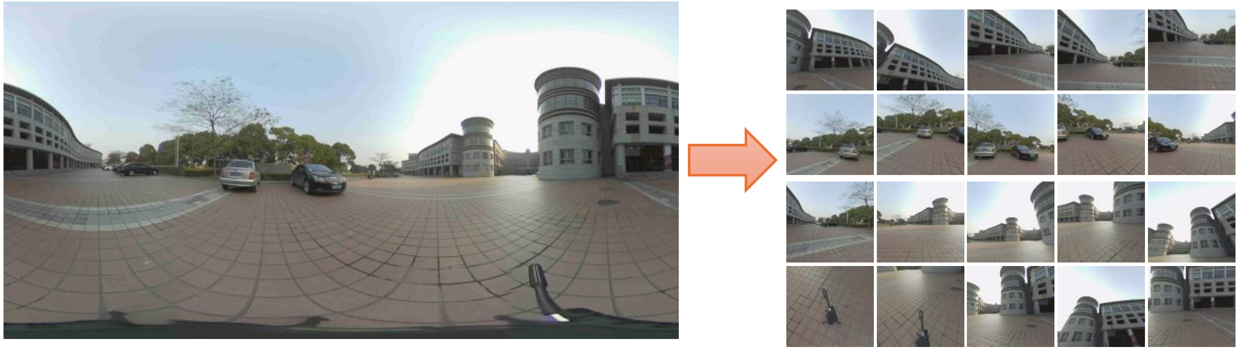


Fig. 5. An example of one entire OI and its twenty viewports based on the work [55].

generated text description can complement visual information and provide us with rich semantic information.

Considering that in human evaluation, ImageReward [52] presents a promising performance for evaluating text-to-image synthesis, we utilize its text encoder, Bert, to obtain the semantic features of the generated text description. For a given text description t, we can obtain the text features through text encoder E_p, as follows:

$$f_t = E_p(t), \tag{1}$$

where f_t represent the text semantic features of text description t.

2) Local Viewports Extraction: Considering that people only watch one local viewport at each time under the HMD, the local visual quality is also crucial for providing quality experience. To simulate human visual perceptual mechanism, 20 viewports are captured following the work [55]. Fig. 5 presents an example of an entire OI and its viewports.

B. Cognition Fusion

Inspired by the declarative and non-declarative cognitive mechanism in human brain [23], [31], the reasonable integration of the text and image information can well model the process of human subjective judgment. Since the non-declarative visual information is extracted from global and local aspects, the text semantic features are fused with global visual features and local visual features, respectively. The details are as follows.

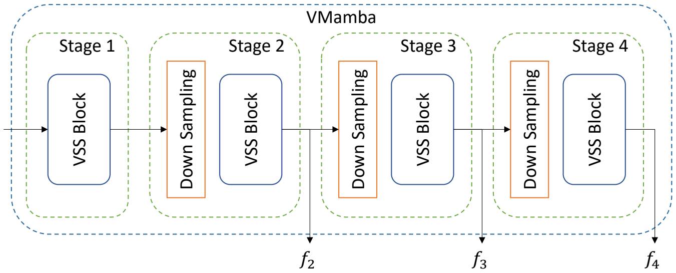


Fig. 6. The three temporary outputs from VMamba [53].

1) Global Feature Fusion (GFF): For the global feature fusion, due to the task insensitivity, the multi-scale features instead of one scale feature are applied to conduct the global multimodal features fusion.

Specifically, the entire OI is first sent into the VMamba module, and the outputs of the last three stages are obtained as multi-scale global visual features. As shown in Fig. 6, given an entire OI, denote the output of the i-th stage of VMamba as f_n, the output of the first stage can be obtained as follows:

$$T_1 = \mathrm{SS2DB}(\mathrm{LN}(M_d)) \oplus M_d, \tag{2}$$

$$f_1 = \mathrm{FFN}(\mathrm{LN}(T_1)) \oplus T_1, \tag{3}$$

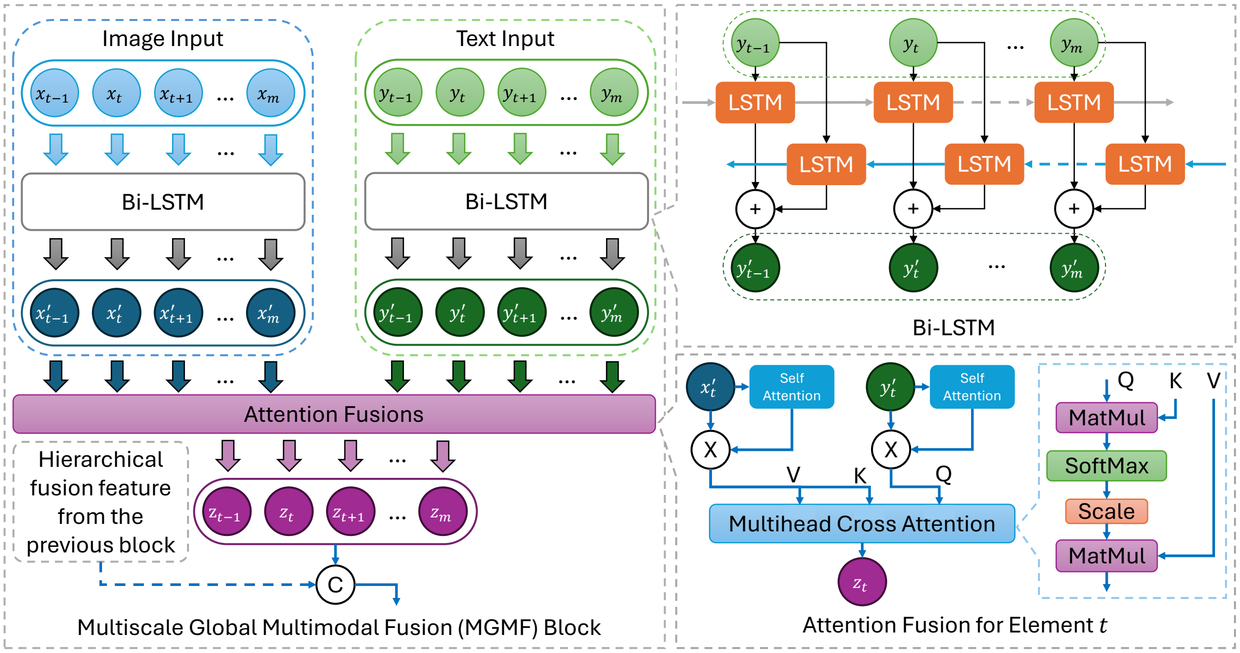


Image Input
x_{t-1} x_t x_{t+1} … x_m
Bi-LSTM
x'_{t-1} x'_t x'_{t+1} … x'_m
Text Input
y_{t-1} y_t y_{t+1} … y_m
Bi-LSTM
y'_{t-1} y'_t y'_{t+1} … y'_m
Attention Fusions
Hierarchical fusion feature from the previous block
z_{t-1} z_t z_{t+1} … z_m
C
Multiscale Global Multimodal Fusion (MGMF) Block
LSTM
+
Bi-LSTM
Self Attention
X
V
K
Q
Multihead Cross Attention
z_t
MatMul
SoftMax
Scale
MatMul
Attention Fusion for Element t

Fig. 7. The diagram of MGMF block and the overview of the attention fusion for the t-th element in two input vectors. This block fuses the multi-scale text semantic features and multi-scale global visual features based on the Bi-LSTM-Attention module.

where M_d is the 2D feature map of OI. SS2DB denotes the SS2D block [53], FFN is a feedforward neural network, and LN means layer normalization. The output of the i-th stage $(i > 1)$ can be obtained as follows:

$$T_i = \text{SS2DB}(\text{LN}(\text{DS}(f_{i-1})) \oplus \text{DS}(f_{i-1}), \tag{4}$$

$$f_i = \text{FFN}(\text{LN}(T_i)) \oplus T_i, \tag{5}$$

where SS2DB denotes the SS2D block [53], FFN is a feedforward neural network, LN means layer normalization, and DS is for down-sampling operation. The outputs f_2, f_3 and f_4 are taken as the multi-scale visual features to conduct the global features fusion. For the declarative text semantic information, the tokens of the text semantic features are sent into a 1-dimensional convolutional layer and three different sizes of convolutional layers to get multi-scale text semantic features.

To better dig into the hidden interactive relationship between text information and global visual information, a multi-scale global multimodal fusion (MGMF) module is designed to fuse the multi-scale text semantic features and multi-scale global visual features, shown in Fig. 7. Specifically, since the entire OI has 360° × 180° wide FoV, image contextual information may be forgotten as the time passes forward. Considering that the Bi-LSTM model has a significant capability on dealing with sequential modeling problems, and the attention mechanisms can enhance model focus on critical information. We introduce the Bi-LSTM-Attention module to strengthen global visual features. Additionally, to well understand the text contextual information, we also adopt the Bi-LSTM-Attention module to strengthen the text semantic features. Then the strengthened text semantic features and visual features are integrated together based on a multihead cross attention to achieve multi-scale interactive fusion from shallow to deep level. To strengthen the hierarchical feature fusion, the previous hierarchical fusion features are concatenated with the current fusion features to increase the representative capability of model. With the incorporation of the above attention mechanisms, the model can effectively capture relationships among different input segments and further refine fusion features, which can reduce the risk of overfitting and improve the overall performance. The whole procedure can be formulated as follows:

$$F_{G1} = \text{AF}(\text{BiLSTM}(\text{Conv}(f_t)), \text{BiLSTM}(f_2)), \tag{6}$$

$$F_{G2} = \text{C}(F_{G1}, \text{AF}(\text{BiLSTM}(\text{Conv}(f_t)), \text{BiLSTM}(f_3))), \tag{7}$$

$$F_{G3} = \text{C}(F_{G2}, \text{AF}(\text{BiLSTM}(\text{Conv}(f_t)), \text{BiLSTM}(f_4))), \tag{8}$$

where f_t is the text semantic feature reshaped by a 1-dimensional convolutional layer, f_2, f_3 and f_4 are the multi-scale visual features. C denotes the concatenation operation. BiLSTM is the Bi-LSTM module, and AF is the attention fusion, which can be formulated as follows:

$$\text{AF}(X, Y) = \text{MCA}(K, Q, V), \tag{9}$$

$$K = V = x_t \otimes \text{SA}(x_t), \tag{10}$$

$$Q = y_t \otimes \text{SA}(y_t), \tag{11}$$

where MCA is the multihead cross attention mechanism as shown in Fig. 7 and SA is the self-attention mechanism. F_{Gi} is the output of the i-th fusion features. In this paper, we take F_{G3} as the global fusion features.

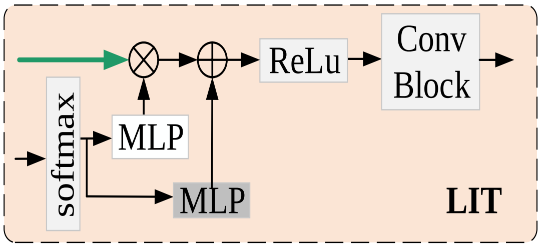


Fig. 8. The Local Image-Text Fusion (LIT) module.

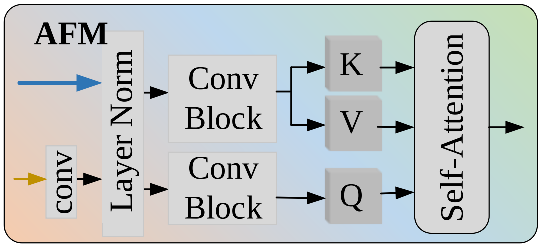


Fig. 9. The Aggregate Fusion Model (AFM) based on attention mechanism.

2) Local Feature Fusion (LFF): To obtain key semantic details and improve the text-image fusion accuracy, local non-declarative visual features and declarative textual semantic information are fine-grained fused, which can not only complement the global fusion features, but also improve the model's ability of understanding semantic information. For the local visual features, ViT [56] is adopted as the image encoder to extract local non-declarative visual features. For a given viewport I_i, we can obtain the visual features of each viewport, as follows:

$$f_{li} = E(I_i), \tag{12}$$

where E is the encoder of ViT, and f_{li} is the local visual features of the i-th viewport.

To achieve effective local feature fusion, we design a local image-text fusion module, named LIT, shown in Fig. 8, to fuse each local viewport features and the text semantic features. Specifically, the text tokens are first reshaped to align with the viewport visual features and multiplied with the viewport visual features passed through the first MLP layer, and added with the viewport visual features passed through the second MLP layer. Then a ReLU layer and a Conv block are introduced to strengthen the fusion, which can improve fusion accuracy and preserve key semantic details. The fusion procedure is shown as follows:

$$\widetilde{f_{ti}} = f_t \otimes \mathrm{MLP}(\mathrm{SoftMax}(f_{li})), \tag{13}$$

$$\widetilde{f_{vti}} = \widetilde{f_{ti}} \oplus \mathrm{MLP}(\mathrm{SoftMax}(f_{li})), \tag{14}$$

$$f_{vti} = \mathrm{Conv}(\mathrm{ReLU}(\widetilde{f_{vti}})), \tag{15}$$

where f_{vti} denotes the local fusion features of each viewport. Then the local fusion features of all viewports are concatenated together to obtain the final local fusion feature, as follows:

$$F_L = \underset{i=1}{\overset{20}{\mathrm{C}}}(f_{vti}), \tag{16}$$

where C denotes the concatenation operation.

C. Quality Prediction

To avoid rigid fusion, we build an aggregate fusion model based on attention mechanism, named AFM, shown in Fig. 9, to ensembled the global fusion features and local fusion features, which can achieve a smooth fusion based on attention mechanism and give a comprehensive quality evaluation. The whole procedure is depicted as follows:

$$K = V = \mathrm{Conv}(\mathrm{LN}(F_{G3})), \tag{17}$$

$$Q = \mathrm{Conv}(\mathrm{LN}(\mathrm{Conv}(F_L))), \tag{18}$$

$$F_A = \mathrm{SelfAttn}(K, Q, V), \tag{19}$$

where F_A denotes the output of the AFM module. SelfAttn denotes the self-attention mechanism.

Finally, we map the above features to the quality score by applying an FC layer. Euclidean loss function is employed to update the parameters of our model, shown as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||s_i - p_i||_2^2, \tag{20}$$

where N is the number of the samples, s_i denotes the ground truth quality score of the i-th sample, and p_i denotes the i-th sample's quality score predicted by the model.

IV. Experimental Results

A. Experimental Settings

1) Datasets: OIQA dataset [57]: It consists of 320 distorted OIs generated from 16 reference images based on two types of compressions (JPEG and JPEG2000) and two types of degradations (Gaussian blur (GB), and Gaussian noise (GN)). Each image is in equireclangular format and has the corresponding Mean Opinion Score (MOS) value.

CVIQ dataset [58]: It includes 528 distorted omnidirectional images and 16 reference images. The distorted omnidirectional images are degraded based on three different types of coding compressions (JPEG [59], H.264/AVC [60] and H.265/HEVC [61]), and each of them has MOS value.

2) Evaluation Criteria: To prove the effectiveness of our model, three criteria are adopted to make the comparison: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Root Mean Squared Error (RMSE), which presents as follows:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2 \sum_{i=1}^{N}(p_i - \bar{p})^2}}, \tag{21}$$

where N denotes the number of samples, s_i and p_i are the ground truth score and the predicted score by the model of the i-th sample, respectively. $\bar{s}$ and $\bar{p}$ are the mean values of s_i and p_i, respectively.

$$\mathrm{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{22}$$

TABLE I
THE PERFORMANCE COMPARISON ON THE OIQA DATASET [57]

	Method	JPEG			JP2K			GN			GB			Overall		
		PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
FR	S-PSNR [64]	0.870	0.829	7.738	0.816	0.849	8.686	0.919	0.885	5.033	0.699	0.692	9.501	0.716	0.712	10.030
	CPP-PSNR [63]	0.865	0.829	7.873	0.849	0.837	7.943	0.920	0.885	5.001	0.672	0.667	9.830	0.707	0.703	10.167
	WS-PSNR [41]	0.861	0.828	7.994	0.844	0.832	8.070	0.922	0.885	4.942	0.661	0.658	9.966	0.689	0.693	10.428
NR	BRISQUE [64]	0.935	0.921	8.689	0.725	0.733	11.355	0.968	0.979	4.551	0.844	0.857	9.161	0.823	0.831	9.262
	DESQUE [65]	0.897	0.868	6.952	0.739	0.732	10.120	0.953	0.937	3.882	0.749	0.663	8.799	0.725	0.712	9.903
	dipIQ [66]	0.829	0.789	8.783	0.916	0.918	6.030	0.955	0.943	3.772	0.932	0.898	4.816	0.701	0.691	10.259
	MEON [67]	0.823	0.779	8.935	0.680	0.601	11.017	0.952	0.930	3.895	0.764	0.716	8.572	0.749	0.717	9.536
	BMPRI [68]	0.918	0.909	6.210	0.185	0.166	14.768	0.961	0.949	3.534	0.356	0.354	12.248	0.431	0.338	12.984
	SSP-BOIQA [15]	0.877	0.834	7.620	0.853	0.852	7.501	0.905	0.843	5.451	0.854	0.862	6.834	0.860	0.865	7.313
	MC360IQA [69]	0.912	0.901	6.535	0.896	0.882	6.573	0.913	0.926	5.240	0.893	0.918	6.072	0.890	0.909	6.697
	VGCN [71]	0.954	0.929	4.288	0.977	0.946	4.313	0.981	0.975	3.617	0.985	0.965	4.213	0.958	0.952	4.385
	PICS (Pro.) [70]	0.968	0.946	3.988	0.980	0.972	4.047	0.989	0.983	3.575	0.990	0.974	3.827	0.970	0.964	3.991
	MMIFN (**ours**)	**0.979**	**0.972**	**3.274**	**0.983**	**0.981**	**3.456**	**0.993**	**0.993**	**3.237**	**0.995**	**0.984**	**3.054**	**0.986**	**0.982**	**3.253**

TABLE II
THE PERFORMANCE COMPARISON ON THE CVIQ DATASET [58]

	Method	JPEG			H.264/AVC			H.265/HEVC			Overall		
		PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
FR	S-PSNR [62]	0.892	0.778	7.727	0.789	0.786	7.589	0.762	0.758	7.785	0.785	0.761	8.714
	CPP-PSNR [63]	0.884	0.765	7.996	0.779	0.777	7.751	0.751	0.748	7.936	0.779	0.754	8.822
	WS-PSNR [41]	0.880	0.756	8.101	0.775	0.773	7.814	0.747	0.744	7.993	0.777	0.751	8.850
NR	BRISQUE [64]	0.913	0.938	5.144	0.780	0.779	7.715	0.771	0.758	8.340	0.826	0.828	7.572
	DESQUE [65]	0.912	0.870	7.003	0.385	0.173	11.410	0.328	0.152	11.362	0.566	0.417	11.603
	dipIQ [66]	0.928	0.793	6.353	0.620	0.635	9.695	0.361	0.326	11.216	0.706	0.623	9.960
	MEON [67]	0.808	0.566	10.057	0.599	0.574	9.900	0.783	0.782	7.484	0.665	0.567	10.510
	BMPRI [68]	0.776	0.498	10.767	0.533	0.520	10.459	0.846	0.840	6.412	0.627	0.621	10.962
	SSP-BOIQA [15]	0.915	0.853	6.847	0.885	0.861	7.042	0.854	0.841	6.302	0.890	0.856	6.941
	MC360IQA [69]	0.941	0.923	5.804	0.932	0.941	5.357	0.914	0.899	4.801	0.939	0.904	4.606
	VGCN [71]	0.989	0.976	2.359	0.972	0.966	3.149	0.940	0.943	4.026	0.965	0.964	3.657
	PICS (Pro.) [70]	0.990	0.983	**2.136**	0.976	0.972	2.967	0.959	0.962	3.577	0.976	0.973	3.290
	MMIFN (**ours**)	**0.992**	**0.990**	2.162	**0.982**	**0.983**	**2.795**	**0.978**	**0.976**	**3.209**	**0.984**	**0.983**	**2.721**

where d_i denotes the distance between the i-th sample's quality score rank in the ground truth scores and the rank in the scores predicted by the model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - p_i)^2}, \tag{23}$$

where s_i and p_i are the ground truth score and the predicted score by the model of the i-th sample, respectively.

Among the three criteria, the higher the PLCC and SROCC values are, the better the performance is. While the lower the RMSE values is, the better the performance is.

3) Implementation Details: In the implementation, we split the dataset into a training set and a testing set with 8:2 proportion. We adopt PyTorch framework to realize our proposed method and execute the fine-tuning operation on both OIQA and CVIQ datasets. The initial learning rate is 1×10^{-3}, and we utilize AdaMax as the optimizer after comparing it with other optimizers such as SGD and Adam. The model is trained and tested on one NVIDIA GeForce RTX 4090 Graphics Card.

B. Comparison with the State-of-The-Arts

In this section, we first compare our model with many state-of-the-arts models on OIQA dataset, and the results are present in Table I, in which the top result is highlighted in boldface. The comparison models can be classified into two categories, FR and NR models. The FR models we adopted here, including S-PSNR [62], CPP-PSNR [63], WS-PSNR [41], are specifically designed for OIQA. Among all the NR models, BRISQUE [64], DESQUE [65], dipIQ [66], MEON [67] and BMPRI [68] is designed for traditional 2D images, while SSP-BOIQA [15], MC360IQA [69], VGCN [70] and PICS (Pro.) [71] are designed for OIQA.

For the overall performance, Table I indicates that the NR models designed for traditional 2D images present a positive overall performance, and even outperforms some FR OIQA models, which indicates that in a certain degree the effective feature extraction is more important. The NR models specially designed for OIQA achieve best overall performance among all the types of models, which indicates that the characteristics of OI itself play a vital role in OIQA field. Among them, PICS (Pro.) model, only taking the visual features into consideration, presents a potential overall performance, while our model achieves the best overall performance on all metrics, which proves the reasonable of combing the text semantic information and visual information. For specific distortions, the proposed model holds the best performance and shows the stability of prediction accuracy and good monotonicity in

TABLE III
THE RESULTS OF THE CROSS-DATASETS VALIDATION

Train	Test	Criterion	BRISQUE	dipIQ	MEON	MC360IQA	MUSIQ	CVRKD-IQA	VGCN	PICS (Pro.)	MMIFN
CVIQ	OIQA	PLCC	0.682	0.583	0.604	0.705	0.762	0.803	0.787	0.827	**0.864**
		SRCC	0.524	0.502	0.551	0.684	0.792	0.801	0.778	0.815	**0.847**
		RMSE	10.870	11.747	11.399	10.178	6.447	6.325	5.437	5.124	**4.866**
OIQA	CVIQ	PLCC	0.754	0.630	0.688	0.823	0.898	0.933	0.924	0.935	**0.941**
		SRCC	0.689	0.587	0.624	0.814	0.901	0.902	0.928	0.931	**0.945**
		RMSE	9.381	10.904	10.145	7.811	5.657	4.892	5.462	4.887	**4.591**

TABLE IV
THE EXPERIMENTAL RESULTS OF THE ABLATION STUDIES

No.	Text	GFF	LFF	AFM	C/AFM	LIT	C/LIT	SA	OIQA			CVIQ		
									PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
(1)	✓	✓	✓	✓	✗	✓	✗	✓	**0.986**	**0.982**	**3.253**	**0.984**	**0.983**	**2.721**
(2)	✓	✓	✓	✓	✗	✓	✗	✗	0.965	0.967	3.332	0.951	0.957	2.987
(3)	✗	✓	✓	✓	✗	✓	✗	✓	0.942	0.935	5.793	0.940	0.931	6.154
(4)	✓	✗	✓	✗	✓	✓	✗	✗	0.658	0.663	9.802	0.672	0.643	9.745
(5)	✓	✓	✗	✗	✓	✓	✗	✓	0.881	0.872	7.841	0.886	0.874	7.753
(6)	✓	✓	✓	✗	✓	✓	✗	✓	0.923	0.924	6.109	0.920	0.919	6.158
(7)	✓	✓	✓	✓	✗	✗	✓	✓	0.951	0.943	4.185	0.950	0.944	4.297

all distortion types, which proves that the proposed model can well predict the OI quality under different types of distortions. Overall, our model can be effectively applied to predict the quality of OI and achieves good consistency with human visual perception.

To further prove the performance of our model, we also compare our model with many state-of-the-arts models on CVIQ dataset, and the results are present in Table II. As we can see, the FR OIQA model achieves better overall performance than the NR models designed for traditional 2D image, and the NR models designed for OIQA still achieve best overall performance among all types of models. Our model achieves the best overall performance among all the models, which proves the reasonable and effectiveness of our model. For specific distortions, our model holds the best performance on most metrics, only RMSE value on JPEG distortion ranks the second. Although PICS (Pro.) model presents the potential performance on RMSE value on JPEG distortion, it fails to predict the quality of other distortions. The problem reason is that the other two distortions are video coding compression methods, which are hard to capture significant quality awareness information [72], [73]. Our model demonstrates its strong ability on these compression distortions, and shows the stability of prediction performance, which demonstrates the reasonable and advantage of our model.

C. Cross-Datasets Validation

To prove the generalization ability and the robustness of the proposed model, we conduct the cross-dataset validation by training the model on one dataset and testing on the other dataset, and the results are listed in Table III. The performances of training on OIQA and testing on CVIQ are better than that of training on CVIQ and testing on OIQA. The problem reason is that OIQA has more complexity types of distortions than CVIQ, which makes predicting OIQA based on training CVIQ challenge. Our model yields the best performance among all the models on these two cross-dataset validation experiments, even on the challenge cross-data experiment (training on CVIQ and testing on OIQA), which proves its stability and generalization ability.

D. Ablation Studies

To prove the contribution of each component of our model, we conduct the ablation experiments, and the results are listed in Table IV. Model (1) is our proposed model, while other models are designed to prove the effectiveness of corresponding components. For example, model (2) is to prove the effectiveness of attention mechanism on Bi-LSTM-Attention module by deleting the self-attention layer, and model (3) is set to prove the effectiveness of text semantic features by deleting the text input. Model (4) means that the global feature fusion module and AFM module are removed, which only takes the local fusion feature to predict the final quality score. While model (5) removes the local feature fusion module and AFM module and takes the global fusion feature to predict the final quality score. Model (6) takes the concatenation operation instead of AFM module to fuse the global fusion features and local fusion features, and model (7) takes the concatenation operation instead of LIT module to fuse the local visual feature and text semantic feature. Model (4) has the worst performance on all the criteria, which proves that global fusion feature plays the most important role compared with other components, and the reasonable of fusing text and visual information. Model (5) ranks second to the last, which proves the reasonable of combing local fusion features, and can improve multimodal feature learning ability. It is one of our contributions. Model (6) demonstrates that the AFM can improve the accuracy of our model, which is also one of our contributions. Model (2) presents the worse performance than the proposed model, which demonstrates that the attention mechanism can significantly improve predictive performance. Overall, the above results prove that our model is an effective and accuracy tool for OIQA.

V. Conclusion

In this paper, we propose a multimodal interactive fusion network (MMIFN) for omnidirectional image quality assessment method. The textual semantic information is firstly introduced to OIQA field, and multimodal interactive fusion between non-declarative visual features and declarative textual semantics are investigated. The global visual features and local visual features are obtained to fuse with textual semantic information, respectively, which can capture rich quality degradation-aware information. In addition, an aggregated fusion module is designed to smoothly fuse the above fusion features. Extensive experiments are conducted to prove the advantage of our model. In the future, we plan to introduce textual information into three-dimensional omnidirectional image quality assessment tasks. Furthermore, considering most of visual tasks are based on the Large Language Model (LLM), the LLM-based OIQA model will be discussed.

Acknowledgments

This work is supported in part by Shenyang science and technology plan project under Grant 23-407-3-32, in part by Natural Science Foundation of Liaoning Province under Grant 2023-MS-139, and in part by National Natural Science Foundation of China under Grant 61901205.

References

[1] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, “Assessing visual quality of omnidirectional videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3516–3530, 2019.

[2] X. Sui, K. Ma, Y. Yao, and Y. Fang, “Perceptual quality assessment of omnidirectional images as moving camera videos,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3022–3034, 2022.

[3] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, 2003.

[4] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] H. Sheikh and A. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[6] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, “Fourier transform-based scalable image quality measure,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3364–3377, 2012.

[7] J. Wu, W. Lin, G. Shi, and A. Liu, “Reduced-reference image quality assessment with visual information fidelity,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1700–1705, 2013.

[8] C. Li, Z. Zhang, H. Wu, K. Zhang, L. Bai, X. Liu, G. Zhai, and W. Lin, “Paps-ovqa: Projection-aware patch sampling for omnidirectional video quality assessment,” in *Proceedings of the 2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2024.

[9] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, “Perceptual video quality assessment: a survey,” *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.

[10] L. Zhang, Y. Shen, and H. Li, “Vsi: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[11] G. Yue, C. Hou, T. Zhou, and X. Zhang, “Effective and efficient blind quality evaluator for contrast distorted images,” *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733–2741, 2019.

[12] Q. Jiang, W. Zhou, X. Chai, G. Yue, F. Shao, and Z. Chen, “A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9784–9796, 2020.

[13] X. Zhu, H. Duan, Y. Cao, Y. Zhu, Y. Zhu, J. Liu, L. Chen, X. Min, and G. Zhai, “Perceptual quality assessment of omnidirectional audio-visual signals,” in *Proceedings of the Artificial Intelligence: Third CAAI International Conference, CICAI 2023, Fuzhou, China, July 22–23, 2023, Revised Selected Papers, Part II*, pp. 512—525, 2024.

[14] K. Liu, T. Sun, H. Zeng, Y. Zhang, C.-M. Pun, and C.-M. Vong, “Spatial-aware conformal prediction for trustworthy hyperspectral image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[15] X. Zheng, G. Jiang, M. Yu, and H. Jiang, “Segmented spherical projection-based blind omnidirectional image quality assessment,” *IEEE Access*, vol. 8, pp. 31 647–31 659, 2020.

[16] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, “Blind quality assessment based on pseudo-reference image,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[17] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[18] Y. Liu, X. Yin, Z. Wan, G. Yue, and Z. Zheng, “Toward a no-reference omnidirectional image quality evaluation by using multi-perceptual features,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 2, 2023.

[19] A. K. R. Poreddy, R. B. C. Ganeswaram, B. Appina, P. Kokil, and R. B. Pachori, “No-reference virtual reality image quality evaluator using global and local natural scene statistics,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[20] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[21] Y. Zhu, Y. Gao, T. Ding, X. Liu, W. Yang, and T. Zhang, “Spatio-temporal pyramid keypoint detection with event cameras,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[22] T. Li, Y. Liu, W. Ren, B. Shiri, and W. Lin, “Single image dehazing using fuzzy region segmentation and haze density decomposition,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[23] C. Yang, X. Han, T. Han, H. Han, B. Zhao, and Q. Wang, “Edge approximation text detector,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[24] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[25] C. Zhang, and S. Liu, “No-reference omnidirectional image quality assessment based on joint network,” in *2022 the 30th ACM International Conference on Multimedia*, pp. 943—951, 2022.

[26] Y. Zhang, L. Wan, D. Liu, X. Zhou, P. An, and C. Shan, “Saliency-guided no-reference omnidirectional image quality assessment via scene content perceiving,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024.

[27] X. Zheng, B. Wang, X. Du, and X. Lu, “Mutual attention inception network for remote sensing visual question answering,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[28] H. Miao, Y. Zhang, D. Wang, and S. Feng, “Multi-output learning based on multimodal gcn and co-attention for image aesthetics and emotion analysis,” *Mathematics*, vol. 9,pp. 1437, 2021.

[29] X. Zhang, X. Gao, L. He, and W. Lu, “Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks,” *Neurocomputing*, vol. 430, pp. 14–23, 2021.

[30] L. Li, T. Zhu, P. Chen, Y. Yang, Y. Li, and W. Lin, “Image aesthetics assessment with attribute-assisted multimodal memory network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7413–7424, 2023.

[31] Z. Zhang, H. Wu, Y. Zhou, C. Li, W. Sun, C. Chen, X. Min, X. Liu, W. Lin, and G. Zhai, “Lmm-pcqa: Assisting point cloud quality assessment with lmm,” 2024.

[32] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, “Smil: Multimodal learning with severely missing modality,” 2021.

[33] D. Kim and T. Kim, “Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models,” 2024.

[34] H. Duan, X. Zhu, Y. Zhu, X. Min, and G. Zhai, “A quick review of human perception in immersive media,” *IEEE Open Journal on Immersive Displays*, vol. 1, pp. 41–50, 2024.

[35] M. Shadlen and D. Shohamy, “Decision making and sequential sampling from memory,” *Neuron*, vol. 90, no. 5, pp. 927–939, 2016.

[36] N. J. Cohen and L. R. Squire, “Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that,” *Science*, vol. 210, no. 4466, pp. 207–210, 1980.

[37] Z. You, Z. Li, J. Gu, Z. Yin, T. Xue, and C. Dong, “Depicting beyond scores: Advancing image quality assessment through multimodal language models,” 2024.

[38] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36, 2015.

[39] V. Zakharchenko, K. P. Choi, and J. H. Park, “Quality metric for spherical panoramic video,” in *Optics and Photonics for Information Processing X*, vol. 9970, International Society for Optics and Photonics, SPIE, pp. 99700C, 2016.

[40] Y. Zhou, M. Yu, H. Ma, H. Shao, and G. Jiang, “Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video,” in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 54–57, 2018.

[41] Y. Sun, A. Lu, and L. Yu, “Weighted-to-spherically-uniform quality evaluation for omnidirectional video,” *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[42] H. G. Kim, H.-T. Lim, and Y. M. Ro, “Deep virtual reality image quality assessment with human perception guider for omnidirectional image,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.

[43] P. Ma, L. Liu, C. Xiao, and D. Xu, “Omnidirectional image quality assessment with mutual distillation,” in *IEEE Transactions on Broadcasting*, vol. 71, no. 1, pp. 264–276, 2025

[44] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[45] H. Jiang, G. Jiang, M. Yu, T. Luo, and H. Xu, “Multi-angle projection based blind omnidirectional image quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4211–4223, 2022.

[46] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, “Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment,” *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 396–405, 2023.

[47] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, “Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1767–1777, 2022.

[48] X. Liu, S. Qiu, M. Zhou, W. Le, Q. Li, and Y. Wang, “Wfanet-ddcl: wavelet-based frequency attention network and dual domain consistency learning for 7t mri synthesis from 3t mri,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[49] L. Tran, X. Liu, J. Zhou, and R. Jin, “Missing modalities imputation via cascaded residual autoencoder,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2017.

[50] X. Zhang, Y. Xiao, J. Peng, X. Gao, and B. Hu, “Confidence-based dynamic cross-modal memory network for image aesthetic assessment,” *Pattern Recognition*, vol. 149, pp. 110227, 2024.

[51] Y. Yin, X. Liu, and Z. Zhang, “Sma-mvs: Segmentation-guided multi-scale anchor deformation patch multi-view stereo,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[52] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, “ImageReward: learning and evaluating human preferences for text-to-image generation,” 2023.

[53] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, “VMamba: visual state space model,” 2024.

[54] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” 2015.

[55] J. Lian, J. Zhang, S. Du, Q. Liu, and J. Liu, “Adversarial diffusion network for dunhuang mural inpainting,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023.

[57] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, “Perceptual quality assessment of omnidirectional images,” in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2018.

[58] W. Sun, W. Luo, X. Min, G. Zhai, X. Yang, K. Gu, and S. Ma, “Mc360iqa: The multi-channel cnn for blind 360-degree image quality assessment,” in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019.

[59] G. Wallace, “The jpeg still picture compression standard,” *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. 18–34, 1992.

[60] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the h.264/avc video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[61] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (hevc) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[62] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36, 2015.

[63] V. Zakharchenko, K. P. Choi, and J. H. Park, “Quality metric for spherical panoramic video,” in *Optics and Photonics for Information Processing X*, vol. 9970, International Society for Optics and Photonics, SPIE, pp. 99700C, 2016.

[64] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[65] Y. Zhang and D. M. Chandler, “An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes,” in *Image Quality and System Performance X*, vol. 8653, International Society for Optics and Photonics, SPIE, pp. 86530J, 2013.

[66] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, “dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs,” *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.

[67] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, “End-to-end blind image quality assessment using deep neural networks,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[68] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, “Blind image quality estimation via distortion aggravation,” *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[69] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, “Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2020.

[70] Y. Zhou, Y. Ding, Y. Sun, L. Li, J. Wu, and X. Gao, “Perceptual information completion-based siamese omnidirectional image quality assessment network,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[71] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[72] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36, 2015.

[73] C. Wang, P. Hu, H. Zhao, Y. Guo, J. Gu, X. Dong, J. Han, H. Xu, and X. Liang, “Uniadapter: All-in-one control for flexible video generation,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

Yun Liu (Member, IEEE) received the Ph.D. degree in communication and information engineering from Tianjin University, China, in 2016. From 2014 to 2015, she was a visiting Ph.D. student at the Visual Space Perception Laboratory, University of California, Berkeley, United States. She is currently an associate professor at the Faculty of Information, Liaoning University, Shenyang, China. Her research interests include multimedia quality assessment, image processing, computer vision, and pattern recognition.

Sifan Li is currently pursuing the M.S. degree in computer science and technology at the Faculty of Information, Liaoning University, Shenyang, China. His research interests include multimedia quality assessment, image processing, efficient training and inference, and computer vision.

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2024. He is currently a Postdoctoral Fellow at Shanghai Jiao Tong University. From Sept. 2019 to Sept. 2020, he was a visiting Ph.D. student at the Schepens Eye Research Institute, Harvard Medical School, Boston, USA. He received the Best Paper Award of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) in 2022. His research interests include perceptual quality assessment, quality of experience, visual attention modeling, extended reality (XR), and multimedia signal processing.

Daoxin Fan is currently pursuing the M.S. degree in computer science and technology at the Faculty of Information, Liaoning University, Shenyang, China. His research interests include multimedia quality assessment, image processing, and computer vision.

Guangtao Zhai (Fellow, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.