MoE-AGIQA: Mixture-of-Experts Boosted Visual Perception-Driven and Semantic-Aware Quality Assessment for AI-Generated Images

Junfeng Yang[1,3], Jing Fu[1,3], Wei Zhang[2*], Wenzhi Cao[1,3*], Limei Liu[1,3], Han Peng[1,3]
[1]Hunan University of Technology and Business
[2]ByteDance
[3]Xiangjiang Laboratory

Abstract

Recently, there has been a surge of interest in AI-Generated Image Quality Assessment (AGIQA). Unlike images in common image quality assessment tasks, AI-generated images may suffer from some unique degradations. To this end, we propose a novel mixture-of-experts boosted visual perception-driven and semantic-aware quality assessment for AI-generated images (MoE-AGIQA). Firstly, we design a visual degradation-aware network to ascertain perceptual rules by emulating human perception of visual degradation. To enhance the diversity of visual degradation-aware features, we additionally devise a prior knowledge injection module, which is pre-trained on specific natural images. Secondly, we devise a semantic-aware network to assess the inconsistency between input text prompts and AI-generated images, and further detect potential semantic problems. Thirdly, we propose to conduct cross-attention on visual degradation-aware and semantic-aware features, so that we can obtain comprehensive quality-aware features and the inherent correlation between these features. Finally, we propose a mixture-of-experts module, involving multiple experts working collaboratively. Each expert is responsible for a specific set of features and outputs a corresponding prediction score. The mixture of multiple experts will ultimately yield a holistic, perceptual quality score. Experimental results on benchmark AGIQA datasets and the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 1 Image Challenge demonstrate our superior performance. The source code is available at https://github.com/37s/MoE-AGIQA.

1. Introduction

With the advent of the Artificial Intelligence Generated Content (AIGC) era, millions of AI-generated images are being created daily using AIGC models, including DALLE [24], Stable Diffusion [27], *etc*. As a crucial indicator, image quality can assist in evaluating the accuracy of these AIGC models, enabling iterative improvements in their performance to produce high-quality AI-generated images, thereby better meeting user needs and expectations. However, unlike images in common image quality assessment tasks [8, 9, 14, 34, 43], AI-generated images may suffer from some unique degradations [32, 44], such as unrealistic structures, unreasonable combinations, and mismatched image-text pairs, *etc*., as depicted in Fig. 1. Therefore, there

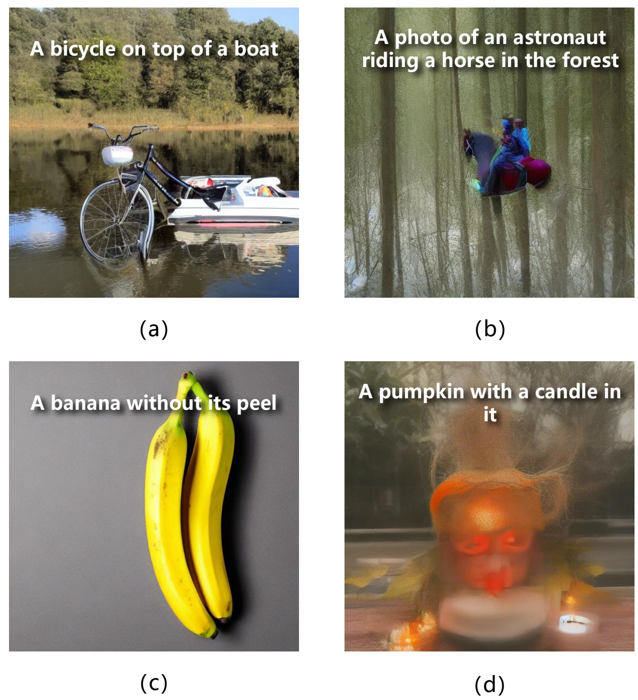

Figure 1. Illustration of some unique degradations of AI-generated images. (a), (b), (c), and (d) are examples of unreasonable combinations, unrealistic structures, mismatched image-text pairs, and AI artifacts, respectively.

*Equal Contribution, Correspondence to Wei Zhang (zhangwei.666@bytedance.com) & Wenzhi Cao (wenzcao@hutb.edu.cn).

is an urgent need to design objective quality assessment models specifically for AI-generated images.

Over the last few years, considerable efforts have been invested in advancing the development of AI-Generated Image Quality Assessment (AGIQA), including the construction of AGIQA datasets like AGIQA-1K [44], AGIQA-3K [15], and AIGCIQA2023 [32], *etc*. Additionally, some AGIQA methods such as PSCR [38] and Q-Align [33], have been proposed to assess the quality of AI-generated images.

Unfortunately, most existing methods [15, 32, 33, 38, 39, 44] predict quality scores based solely on AI-generated images, without considering the text prompts of these images. This significantly limits the effectiveness of these methods in coping with the problem of inconsistency between images and texts. Furthermore, some methods such as TIER [40] cannot correlate well with human visual perception of AI-generated images even when taking the information of text prompts into account.

In summary, we identify three key challenges for evaluating AI-generated images. First, conventional Image Quality Assessment (IQA) methods are primarily designed for natural images. How can we accommodate existing methods for AI-generated images? Second, how can we verify whether AI-generated images are correlated with text prompts, and evaluate image quality from a human visual perception perspective simultaneously? Third, once we have collected efficient features, how can we output a final score that not only simulates the human decision-making process, but also emulates the subjective score accurately?

To tackle the challenges mentioned above, we propose a novel Mixture-of-Experts (MoE) boosted AGIQA model (MoE-AGIQA) that is visual perception-driven and semantic-aware. We start by introducing a visual degradation-aware network to ascertain perceptual rules by emulating human perception of visual degradation. To enhance the diversity of visual degradation-aware features, we further devise a prior knowledge injection module, which is pre-trained on specific natural images. Meanwhile, we devise a semantic-aware network to assess the inconsistency between input text prompts and AI-generated images, and further detect potential semantic problems. To obtain comprehensive quality-aware features and the inherent correlation between visual degradation-aware features and semantic-aware features, we conduct a cross-attention fusion strategy on these features. Inspired by the MoE framework [13, 29], we propose a Top-K expert module that dynamically selects experts for quality prediction, facilitating adaptive learning of degradation-specific knowledge. The main contributions are summarized as follows:

- We present a novel MoE-boosted AGIQA model (MoE-AGIQA), which evaluates the quality of AI-generated images in a visual perception-driven and semantic-aware manner.
- We propose to design IQA features for AI-generated images from a human visual perception perspective, where we devise a visual degradation-aware network, a semantic-aware network, and a natural degradation priors injection module to enrich the diversity of visual quality-aware features.
- We propose a Top-K expert quality prediction module, adaptively and comprehensively computing quality scores for AI-generated images. Extensive experiments on benchmark AGIQA datasets demonstrate that our method outperforms the state-of-the-art.

2. Related Work

2.1. Image Quality Assessment

The purpose of IQA is to automatically predict the quality of images, mimicking the perceptual preferences of human observers. In the last decades, numerous IQA methods have been proposed. Despite significant successes they have achieved in assessing common images (*e.g.* , natural, graphic, and screen content images) [6, 21, 30, 37, 42, 43], IQA for AI-generated images remains a challenge. As a new branch of IQA, there is a relative lack of research on AGIQA. Yuan *et al*. [38] propose a patches sampling-based contrastive regression framework, named PSCR, to leverage differences among various AI-generated images for enhancing representation learning. Despite having overcome the limitations of previous models in utilizing reference images on a no-reference image dataset, they still struggle to address the issue of image-text mismatch, as they rely solely on AI-generated images for quality assessment. To address this issue, Yuan *et al*. [40] propose a text-image encoder-based regression framework, called TIER, which uses an image encoder and a text encoder to extract features from the AI-generated images and corresponding text prompts, respectively.

Different from the methods discussed previously, our method considers both visual degradation and semantic information, and uses natural degradation priors to further enhance the representation of visual degradation. Additionally, we obtain comprehensive quality-aware features by implementing a cross-attention mechanism that enables heterogeneous feature fusion. To make reliable quality predictions for AI-generated images, we define multiple experts and select the Top-K experts from them to collaboratively complete the prediction process.

2.2. Vision-Language Model

Vision-language models have garnered significant attention in recent years due to their outstanding performance in multi-modal learning. Among the pioneering models in this field are CLIP [23] and BLIP [17], which have achieved impressive results in various visual understand-

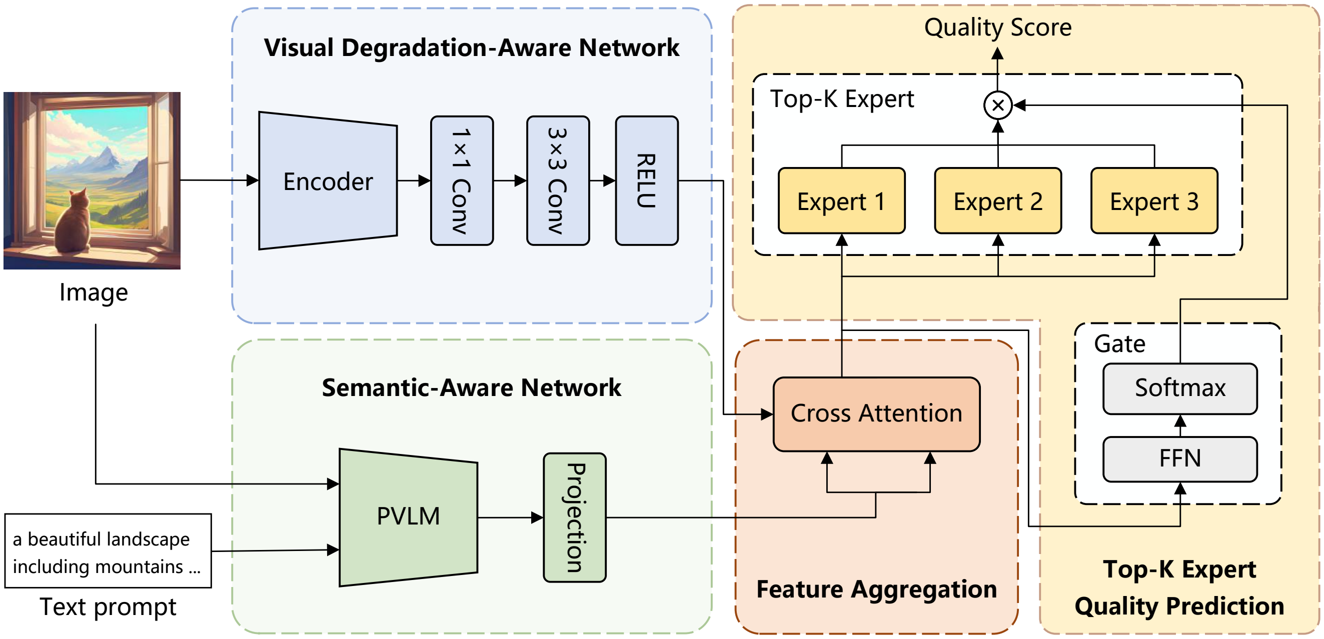

Figure 2. The overall architecture of our model. Given an input AI-generated image, we aim to evaluate its quality. We extract visual degradation-aware features from the AI-generated image. Meanwhile, we use a pre-trained vision-language model (PVLM) with the AI-generated image and its corresponding text prompt, to generate semantic-aware features, and transform them by a linear projection to the same embedding space with visual degradation-aware features. The quality-aware features are obtained through a cross-attention module. Among them, visual degradation-aware features serve as query (Q), and semantic-aware features serve as key (K) and value (V). We define a list of n quality prediction experts. First, the quality-aware features are parsed into the weight scores of experts through the feedforward network (FFN) and a softmax layer, and then the weight scores are used to find the best k experts. Finally, the final quality score is obtained by weighting the quality scores predicted by the best k experts.

ing tasks. This paper proposes leveraging the Pre-trained Vision-Language Model (PVLM) (*e.g.* , [35]) as the backbone of the semantic-aware network to generate features sensitive to semantic content, thus enriching the diversity of quality-aware representation.

2.3. Sparse Mixture of Expert

Sparse MoE [7, 26] is a variant of the MoE [12] framework that emphasizes efficiency and scalability by employing sparsity. In traditional MoE models, all experts contribute to the prediction, which can be computationally expensive, especially when dealing with a large number of experts. In sparse MoE models, only a subset of experts actively participates in the prediction process for a given input, while the remaining experts are dormant. The selection of active experts is typically determined dynamically based on the input data, often through a gating mechanism [3]. This allows the model to bypass unnecessary computations and focus only on the most relevant experts for a given input, leading to improved efficiency and reduced computational costs. We utilize the degradation-specific knowledge in quality-aware features to dynamically select experts and adaptively apply experts to predict the quality scores of AI-generated images.

3. Method

Overview. Given an input image generated by the Text-to-Image (T2I) model, our goal is to predict its quality score in conjunction with its corresponding textual prompt. Our method leverages visual degradation information derived from a pre-trained encoder, and semantic information obtained from a Pre-trained Vision-Language Model (PVLM), allowing for adaptive and comprehensive quality assessment. The overall architecture is illustrated in Fig. 2.

Specifically, we adaptively and comprehensively predict the quality of the input AI-generated image, divided into three steps: i) visual degradation and semantic measurement, using the visual degradation-aware network and the semantic-aware network to individually measure the visual degradation and semantics of the AI-generated image; ii) quality-aware feature aggregation, acquiring quality-aware representation by aggregating visual degradation-aware and semantic-aware features; iii) Top-K expert quality prediction, selecting the best k experts from a candidate list of n quality prediction experts and obtaining the final quality score by weighting the quality scores predicted by these ex-

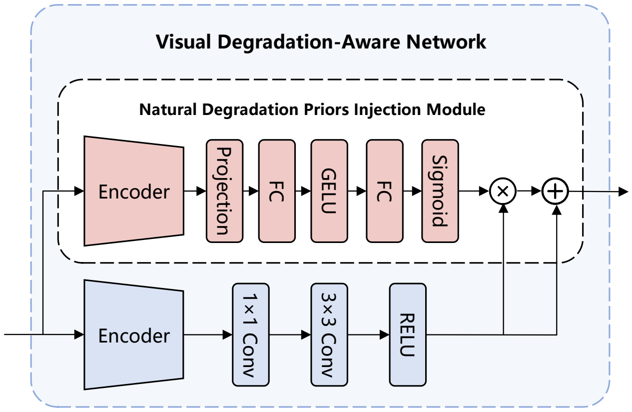

Figure 3. The framework of the visual degradation-aware network.

perts.

3.1. Natural Degradation Prior

We incorporate natural degradation priors to improve the representation of visual degradation in our quality prediction model. In Fig. 3, natural degradation prior knowledge H^{prior} is obtained from a pre-trained encoder (*e.g.* , [5]), and the process can be expressed as:

$$H^{prior^{v1}} = Enc_p(I^d), H^{prior^{v1}} \in \mathbb{R}^{1\times C^{v1}} \tag{1}$$

where $Enc_p(\cdot)$ represents the pre-trained encoder. The class token state at the output of [5] is denoted as $H^{prior^{v1}}$, with C^{v1} representing the channel dimension. Additionally, to ensure that $H^{prior^{v1}}$ aligns with the embedding space of size C for our quality prediction model, we employ a multi-layer perceptron network $MLP(\cdot)$, which can be formulated as:

$$H^{prior} = MLP(H^{prior^{v1}}), H^{prior} \in \mathbb{R}^{1\times C} \tag{2}$$

3.2. MoE Boosted AGIQA Model

Visual Degradation and Semantic Measurement. To effectively deal with various degradations in AI-generated images, we comprehensively consider learning both visual degradation and semantic information. Specifically, we measure the visual degradation-aware feature map M^v and semantic-aware feature map M^s for the AI-generated image I^d,

$$M^{v^{v1}} = Enc_v(I^d), M^{v^{v1}} \in \mathbb{R}^{L^1\times 4C^{v1}} \tag{3}$$

$$M^{s^{v1}} = PVLM(I^d, P^{txt}), M^{s^{v1}} \in \mathbb{R}^{L^2\times C^{v1}} \tag{4}$$

where $Enc_v(\cdot)$ is the pre-trained encoder (*e.g.* , [5]), $PVLM(\cdot,\cdot)$ represents the PVLM, L^1 denotes the number of patches for I^d, and L^2 indicates the sequence length of text prompt P^{txt}. The $M^{v^{v1}}$ is then passed to a feed-forward network $FFN_v(\cdot)$ for improving the feature locality.

$$M^v = FFN_v(M^{v^{v1}}), M^v \in \mathbb{R}^{L^1\times C} \tag{5}$$

Meanwhile, the $M^{s^{v1}}$ transformed by a linear projection $Proj(\cdot)$ to the same embedding space of size C as M^v.

$$M^s = Proj(M^{s^{v1}}), M^s \in \mathbb{R}^{L^2\times C} \tag{6}$$

Quality-Aware Feature Aggregation. Since the visual degradation-aware and semantic-aware features are heterogeneous, we utilize a cross-attention mechanism to enable the heterogeneous fusion of these features. The process is described as:

$$Q = W^q(Norm(M^v)) \tag{7}$$

$$K, V = W^k(Norm(M^s)), W^v(Norm(M^s)) \tag{8}$$

$$F^{d^{v1}} = Attention(Q, K, V) = Softmax\left(Q, K^\top\right)V \tag{9}$$

$$F^d = Norm(M^v) + F^{d^{v1}} \tag{10}$$

where $Norm(\cdot)$ is LayerNorm, W^q, W^k, W^v are linear projection functions. After passing through the cross-attention fusion module, we obtain the comprehensive quality-aware features F^d.

Top-K Expert Quality Prediction. The degradation of AI-generated images varies significantly, making it challenging to predict their quality consistently. To address this issue, we utilize the degradation-specific knowledge in quality-aware representation for adaptive and comprehensive quality assessment. We have n candidate quality prediction experts, $\{E_i|i = 1, ..., n\}$, tasked with handling different AI-generated images containing various degradation types. Each candidate expert E_i specializes in mapping distortion representations of specific degradation types to quality scores. Specifically, the quality-aware features F^d are employed as the input for a feedforward network FFN^t, followed by a Softmax function that outputs normalized selection scores $W^{selection}$ for n candidate experts,

$$W^{selection} = Softmax(FFN^t\left(F^d\right)) \tag{11}$$

The set $\left\{W_i^{selection}|i = 1, ..., n\right\}$ represents the likelihood of utilizing the i-th expert E_i to map the distortion representation of I^d. To obtain a more reliable quality score, we opt for the Top-K experts to make predictions collectively.

Consequently, the final quality score S is computed by assigning weights to the quality scores $(S_j, \ldots, S_k)$ predicted by the best k experts.

$$S = \sum_{j=1}^{k} \left(S_j \cdot W_j^{selection} \right) \tag{12}$$

4. Experiment

4.1. Datasets

Pre-training Dataset. We use the IQA dataset KonIQ-10K [11] to capture the human visual perception of realistic distortions. Specifically, KonIQ-10K comprises 10,073 natural images, which are selected from a large-scale multimedia dataset named YFCC100M [31].

Evaluation Datasets. Our method is evaluated on three publicly available AGIQA datasets, including AGIQA-1K [44], AGIQA-3K [15], and AIGCIQA2023 [32]. AGIQA-1K consists of 1,080 AI-generated images produced by two T2I models stable-inpainting-v1 and stable-diffusion-v2 [27]. AGIQA-3K is the largest among the three AGIQA datasets, which contains 2,982 images generated from six T2I models including four diffusion-based models (GLIDE [22], Stable Diffusion V-1.5 [27], Stable Diffusion XL-2.2 [28], Midjourney [10]), one GAN-based model (AttnGAN [36]), and one auto-regressive-based model (DALLE2 [25]). AIGCIQA2023 consists of 2,400 AI-generated images created by six T2I models (such as Lafite [45], Unidiffuser [1], and Controlnet [41], *etc.*) based on 100 text prompts. For each AGIQA dataset, 80% of the AI-generated images contained in it are randomly sampled for training and the rest 20% are used for testing.

4.2. Evaluation Metrics

Spearman's Rank-Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) are employed as evaluation metrics for our method, measuring prediction monotonicity and precision, respectively. Both SRCC and PLCC range from 0 to 1, with higher values indicating a better performance of the AGIQA method. Furthermore, a comprehensive metric known as the main score, derived from the mean average of PLCC and SRCC, is also provided.

4.3. Implementations Details

Our model has two generations, named MoE-AGIQA-v1 and MoE-AGIQA-v2, respectively. Among them, MoE-AGIQA-v1 is used in NTIRE 2024 Quality Assessment for AI-Generated Content - Track 1 Image Challenge, and MoE-AGIQA-v2 is an optimized version of MoE-AGIQA-v1 that introduces natural degradation priors. Specifically, the only difference between them lies in the presence of the

Table 1. Quantitative comparison on the AGIQA-1K dataset. The best and the second-best performance results are marked in boldface and italics, respectively.

Method	AGIQA-1K SRCC	AGIQA-1K PLCC	AGIQA-1K Main Score
ResNet50 [44]	0.6365	0.7323	0.6844
StairIQA [44]	0.5504	0.6088	0.5796
MGQA [44]	0.6011	0.6760	0.6386
WaDIQaM-NR [2]	0.7280	0.7791	0.7536
CONTRIQUE [20]	0.7930	0.8583	0.8257
PSCR [38]	0.8430	0.8403	0.8417
TIER [40]	0.8266	0.8297	0.8282
MoE-AGIQA-v1	**0.8530**	*0.8877*	*0.8704*
MoE-AGIQA-v2	*0.8501*	**0.8922**	**0.8712**

Table 2. Quantitative comparison on the AGIQA-3K dataset. The best and the second-best performance results are marked in boldface and italics, respectively.

Method	AGIQA-3K SRCC	AGIQA-3K PLCC	AGIQA-3K Main Score
DBCNN [15]	0.8207	0.8759	0.8483
CLIPIQA [15]	0.8426	0.8053	0.8240
CNNIQA [15]	0.7478	0.8469	0.7824
WaDIQaM-NR [2]	0.2187	0.3934	0.3061
CONTRIQUE [20]	0.8073	0.8866	0.8470
PSCR [38]	0.8498	0.9059	0.8779
TIER [40]	0.8251	0.8821	0.8536
MoE-AGIQA-v1	**0.8758**	**0.9294**	**0.9026**
MoE-AGIQA-v2	*0.8746*	*0.9282*	*0.9014*

natural degradation priors injection module. MoE-AGIQA-v1 lacks this module, while MoE-AGIQA-v2 incorporates it. Our experiments are all implemented using PyTorch 2.0.0 and CUDA 12.0 based on a PC with four NVIDIA A100 Tensor Core GPUs.

Pre-training. We utilize a ViT-Base/16 [5] with a two-layer MLP as the original architecture of the natural degradation priors injection module, which is pre-trained on KonIQ-10K. The batch size is set to 16. We use the AdamW [19] optimizer, with a weight decay of 1×10^{-5}, a learning rate of 1×10^{-5} and a cosine annealing scheduler. It takes about 10 hours to train the natural degradation priors injection module, for 200 epochs.

Fine-tuning. During training, our model is trained for 100 epochs with a batch size of 16. The AdamW optimizer with a weight decay of 1×10^{-5} is employed. The learn-

Figure 4. Quality prediction ability of our method on AI-generated images produced by unseen T2I models in cross-dataset experiments. We test MoE-AGIQA-v2, trained on AGIQA-3K, using the full AIGCIQA2023 dataset. Specifically, we assess eight AI-generated images produced by two unseen T2I models (Lafite and Unidiffuser) from the AIGCIQA2023 dataset. The predicted quality scores generated by MoE-AGIQA-v2 and MOS scores (higher is better) are placed at the bottom of each AI-generated image. Remarkably, both the rankings of predicted quality scores and subjective MOS scores are identical.

Table 3. Quantitative comparison on the AIGCIQA2023 dataset. The best and the second-best performance results are marked in boldface and italics, respectively.

Method	AIGCIQA2023		
	SRCC	PLCC	Main Score
CNNIQA [32]	0.7160	0.7937	0.7549
VGG16 [32]	0.7961	0.7973	0.7967
VGG19 [32]	0.7733	0.8402	0.8068
ResNet18 [32]	0.7583	0.7763	0.7673
ResNet34 [32]	0.7229	0.7578	0.7404
WaDIQaM-NR [2]	-	-	-
CONTRIQUE [20]	0.8048	0.8271	0.8160
PSCR [38]	0.8371	0.8858	0.8615
TIER [40]	0.8194	0.8359	0.8277
MoE-AGIQA-v1	*0.8729*	*0.8860*	*0.8795*
MoE-AGIQA-v2	**0.8751**	**0.8904**	**0.8828**

ing rate is initialized with 1×10^{-5} and scheduled by the cosine annealing strategy. Since we use ViT-Base/16 [5] pre-trained on ImageNet [4] as the backbone of the visual degradation-aware network, we random crop all input images into three sub-images with a spatial size of 224×224 or 384×384. For the sake of computational efficiency, we use 224×224 as the size of the input image in our experiments. Meanwhile, for the semantic-aware network using the pre-trained ImageReward [35] as the backbone, we resize all input images to 224×224. Moreover, the backbone of the visual degradation-aware network is frozen, 50% of the transformer layers in the backbone of the semantic-aware network are frozen, and the parameters of the remaining modules can be tunable. The training loss applied is the mean absolute error loss. During testing, for the visual degradation-aware network, each input image is randomly cropped 15 times. The final quality score is computed as the mean of the quality scores from each cropped sub-image.

4.4. Results

Quantitative Comparison. We conduct comparisons with existing methods on three AGIQA datasets and present the performance results in Tab. 1, Tab. 2, and Tab. 3, respectively. Our method achieves state-of-the-art performance. Based on the results, we can draw several conclusions. Firstly, our method benefits from the pair-wise learning strategy, allowing it to acquire both visual degradation-aware and semantic-aware information. As a result, it outperforms purely image-driven methods such as DBCNN [15], CNNIQA [15], and PSCR [38], *etc*. Secondly, our method dynamically selects sparse experts to learn shared and distortion-specific knowledge. By leveraging this adaptive learning mechanism, our method is able to efficiently identify and utilize relevant expertise for different degrees and types of degradation present in AI-generated images. Furthermore, by introducing prior knowledge of reality distortions, the performance of our method is further improved

Table 4. The SRCC and PLCC results of MoE-AIGIQA-v2 on cross-dataset experiments. The best performance results are marked in boldface.

Train	Test	WaDIQaM-NR SRCC	WaDIQaM-NR PLCC	CONTRIQUE SRCC	CONTRIQUE PLCC	MoE-AGIQA-v2 SRCC	MoE-AGIQA-v2 PLCC
AGIQA-1K	AGIQA-3K	-0.0672	-0.978	0.3024	0.2925	**0.7483**	**0.7506**
	AIGCIQA2023	0.1472	0.1463	0.1113	0.1436	**0.4295**	**0.4440**
AGIQA-3K	AGIQA-1K	0.1066	-0.0860	**0.7197**	0.7906	0.7101	**0.8008**
	AIGCIQA2023	0.0136	-0.0207	0.6857	0.7078	**0.7619**	**0.7547**
AIGCIQA2023	AGIQA-1K	-	-	0.5810	0.6798	**0.6619**	**0.7444**
	AGIQA-3K	-	-	0.6372	0.6561	**0.7500**	**0.8152**

Table 5. Comparison of each component of our method on the AIGCIQA2023 dataset.

ViDN	SeAN	NDPM	TKEM	AIGCIQA2023 SRCC	AIGCIQA2023 PLCC
✓	×	×	×	0.8452	0.8654
×	✓	×	×	0.8640	0.8787
✓	✓	×	×	0.8671	0.8808
✓	✓	×	✓	0.8729	0.8860
✓	✓	✓	✓	**0.8751**	**0.8904**

Table 6. Ablation study of different input sizes for ViT in the backbone of the visual degradation-aware network on the AGIQA-3K dataset.

Input Size	AGIQA-3K SRCC	AGIQA-3K PLCC
224×224	0.8746	0.9282
384×384	**0.8789**	**0.9294**

on AGIQA-1K and AIGCIQA2023 datasets. This underscores the advantage of leveraging such pre-existing insights. In conclusion, these findings underscore the effectiveness of our method in addressing the AGIQA task. Our method not only outperforms existing methods but also showcases a promising direction for future research in this field.

Qualitative Comparison. To demonstrate the generalization ability of our method, we conduct cross-dataset experiments. Specifically, WaDIQaM-NR [2] and CONTRIQUE [20] are selected for comparison. The results in Tab. 4 indicate that our method effectively handles a variety of AI-generated images using a single set of parameters. Furthermore, the alignment between the rankings of predicted quality scores and subjective MOS values shown in Fig. 4 further emphasizes the robust generalization ability of our method.

Table 7. Ablation study of various combinations for the outputs of different layers of ViT in the backbone of the visual degradation-aware network on the AIGCIQA2023 dataset.

	AIGCIQA2023 SRCC	AIGCIQA2023 PLCC
$l = 12$	0.8677	0.8853
$l = 1, 2, 3, 4$	0.8656	0.8861
$l = 5, 6, 7, 8$	**0.8751**	**0.8904**
$l = 9, 10, 11, 12$	0.8661	0.8853

4.5. Ablation Studies

Model Architecture. In Tab. 5, we provide an ablation study to verify the effectiveness of the visual degradation-aware network (ViDN), semantic-aware network (SeAN), natural degradation priors injection module (NDPM), and Top-K expert quality prediction module (TKEM). The results indicate that each component plays a crucial role in achieving optimal performance.

Visual Degradation-Aware Network. From Tab. 6, it is evident that utilizing ViT-Base/16 with an input size of 384×384 yields optimal performance. This indicates a bigger resolution image provides more space to capture richer visual degradation representations. In addition, in Tab. 7, we test the outputs of different layers of ViT on the AIGCIQA2023 dataset. Our model performs best when selecting the outputs of the 5th, 6th, 7th, and 8th layers.

Semantic-Aware Network. The semantic-aware network is proposed to acquire semantic information. By introducing this network, we can observe a significant improvement in the performance of SRCC and PLCC. This proves the effectiveness of semantic information. Furthermore, in Tab. 8, we test different fixed rates for the transformer layers in the backbone of the semantic-aware network on the

Table 8. Ablation study of different fixed rates for the transformer layers in the backbone of the semantic-aware network on the AIGCIQA2023 dataset.

AIGCIQA2023	0.1	0.3	0.5	0.7
SRCC	0.8685	0.8701	**0.8751**	0.8694
PLCC	0.8879	0.8873	**0.8904**	0.8865

Table 9. Ablation study of different combinations of the number of candidate experts n and the number of selected experts k on the AIGCIQA2023 dataset. Our model performed best when $n = 4$ and $k = 3$.

	AIGCIQA2023	
	SRCC	PLCC
$n = 2, k = 1$	0.8703	0.8880
$n = 2, k = 2$	0.8714	0.8871
$n = 3, k = 1$	0.8613	0.8787
$n = 3, k = 2$	0.8628	0.8797
$n = 3, k = 3$	0.8614	0.8851
$n = 4, k = 1$	0.8633	0.8803
$n = 4, k = 2$	0.8668	0.8858
$n = 4, k = 3$	**0.8751**	**0.8904**
$n = 4, k = 4$	0.8684	0.8866

AIGCIQA2023 dataset. Our model performs best when the fixed rate is set to 0.5.

Natural Degradation Priors Injection Module. The natural degradation priors injection module is proposed to introduce human impressions of realistic distortions. Results in Tab. 5 show that such prior knowledge is essential for our method.

Top-K Expert Quality Prediction Module. Different AI-generated images often exhibit varying degrees and types of degradation and should be adaptively and comprehensively evaluated. Specifically, we select specific combinations of experts for different AI-generated images. In Tab. 5, this module brings performance gains in SRCC and PLCC. Furthermore, in Tab. 9, we tested different combinations of the number of candidate experts n and the number of selected experts k on the AIGCIQA2023 dataset. Our model performed best when $n = 4$ and $k = 3$.

4.6. Results of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 1 Image Challenge

The objective of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 1 Image Challenge [18] is to develop a solution that accurately predicts the quality of AI-generated images produced by T2I models in the AIGIQA-20K dataset [16], thereby fostering advancements in the field of multi-modal generation. The final results of the challenge on the testing data are reported in Tab. 10, where our method achieved sixth place in terms of the main score.

Table 10. Results of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 1 Image Challenge on the AIGIQA-20K dataset. Our method won sixth place in the challenge.

Team	Main Score
1st	0.9175
2nd	0.9169
3rd	0.9157
4th	0.9138
5th	0.9091
MoE-AGIQA-v1 (ours)	**0.9087**
7th	0.9068
8th	0.9044
9th	0.9023
10th	0.8835
11th	0.8736
12th	0.8715
13th	0.8628
14th	0.8613
15th	0.8595

5. Conclusion

We propose a novel MoE-boosted AGIQA model, named MoE-AGIQA, which evaluates the quality of AI-generated images in a visual perception-driven and semantic-aware manner. The key insight is to design features from a human visual perception perspective and emulate the human decision-making process. Specifically, we propose a visual degradation-aware network, a semantic-aware network, and a natural degradation priors injection module to enrich the diversity of visual quality-aware features. We then predict the quality score of AI-generated images with three steps: visual degradation and semantic measurement, quality-aware feature aggregation, and Top-K expert quality prediction. Experiments on benchmark AGIQA datasets show that our method outperforms the state-of-the-art by a large margin.

Acknowledgments

This work was supported in part by the Major Project of Xiangjiang Laboratory under Grants 23XJ01003 and 23XJ01007.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 5

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1): 206–219, 2017. 5, 6, 7

[3] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5, 6

[6] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 2

[7] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems*, 5, 2023. 3

[8] Ke Gu, Shiqi Wang, Guangtao Zhai, Siwei Ma, Xiaokang Yang, Weisi Lin, Wenjun Zhang, and Wen Gao. Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure. *IEEE Transactions on Multimedia*, 18(3):432–443, 2016. 1

[9] Chunle Guo, Ruiqi Wu, Xin Jin, Linghao Han, Weidong Zhang, Zhi Chai, and Chongyi Li. Underwater ranker: Learn which is better and how to be better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 702–709, 2023. 1

[10] David Holz. "midjourney". https://docs.midjourney.com, 2023. 5

[11] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 5

[12] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3

[13] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 2

[14] Chunyi Li, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai. A full-reference quality assessment metric for cartoon images. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2022. 1

[15] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2, 5, 6

[16] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiqa-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2024. 8

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[18] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 8

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[20] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 5, 6, 7

[21] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, and Weisi Lin. Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transactions on Image Processing*, 26(11):5462–5474, 2017. 2

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 5

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. Pmlr, 2021. 1

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image gener-

ation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5

[26] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 5

[28] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 5

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2

[30] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 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*, pages 3667–3676, 2020. 2

[31] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[32] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. 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*, pages 46–57. Springer, 2023. 1, 2, 5, 6

[33] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 2

[34] Tianhe Wu, Shuwei Shi, Haoming Cai, Mingdeng Cao, Jing Xiao, Yinqiang Zheng, and Yujiu Yang. Assessor360: Multi-sequence network for blind omnidirectional image quality assessment. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[35] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6

[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 5

[37] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 2

[38] Jiquan Yuan, Xinyan Cao, Linjing Cao, Jinlong Lin, and Xixin Cao. Pscr: Patches sampling-based contrastive regression for aigc image quality assessment. *arXiv preprint arXiv:2312.05897*, 2023. 2, 5, 6

[39] Jiquan Yuan, Xinyan Cao, Changjin Li, Fanyi Yang, Jinlong Lin, and Xixin Cao. Pku-i2iqa: An image-to-image quality assessment database for ai generated images. *arXiv preprint arXiv:2311.15556*, 2023. 2

[40] Jiquan Yuan, Xinyan Cao, Jinming Che, Qinyuan Wang, Sen Liang, Wei Ren, Jinlong Lin, and Xixin Cao. Tier: Text and image encoder-based regression for aigc image quality assessment. *arXiv preprint arXiv:2401.03854*, 2024. 2, 5, 6

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5

[42] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 2

[43] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 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*, pages 14071–14081, 2023. 1, 2

[44] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 440–445. IEEE, 2023. 1, 2, 5

[45] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 5