SF-IQA: Quality and Similarity Integration for AI Generated Image Quality Assessment

Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen*
University of Science and Technology of China
{yuzihao, guanfb, luyt31415, lixin666}@ mail.ustc.edu.cn
chenzhibo@ustc.edu.cn

Abstract

In recent years, the rapid development of Artificial Intelligence (AI) has facilitated the widespread use of AI-Generated Images (AIGIs), a subset of Artificial Intelligence Generated Content (AIGC). However, there are prevalent issues associated with AIGIs, notably the unsatisfied quality of the generated images and the misalignment between the generated images and their corresponding textual prompts. These challenges underscore the importance of Image Quality Assessment (IQA) in the field of AIGIs to provide more precise quality predictions that are consistent with human perception. Responding to this need, we introduce SF-IQA, a novel AIGC image quality metric that integrates quality and similarity in a score fusion manner. Specifically, we employ a multi-layer feature extractor and fusion module to extract and aggregate the local and global-level features, facilitating the excavation of quality-aware features. For image-text similarity, we finetuned a strong vison-language model based on a powerful perceptual-aware image-text alignment prior. With the assistance of score fusion manner, our proposed SF-IQA obtains state-of-the-art results on AGIQA-3K benchmarks and achieves 4th place in the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge.

1. Introduction

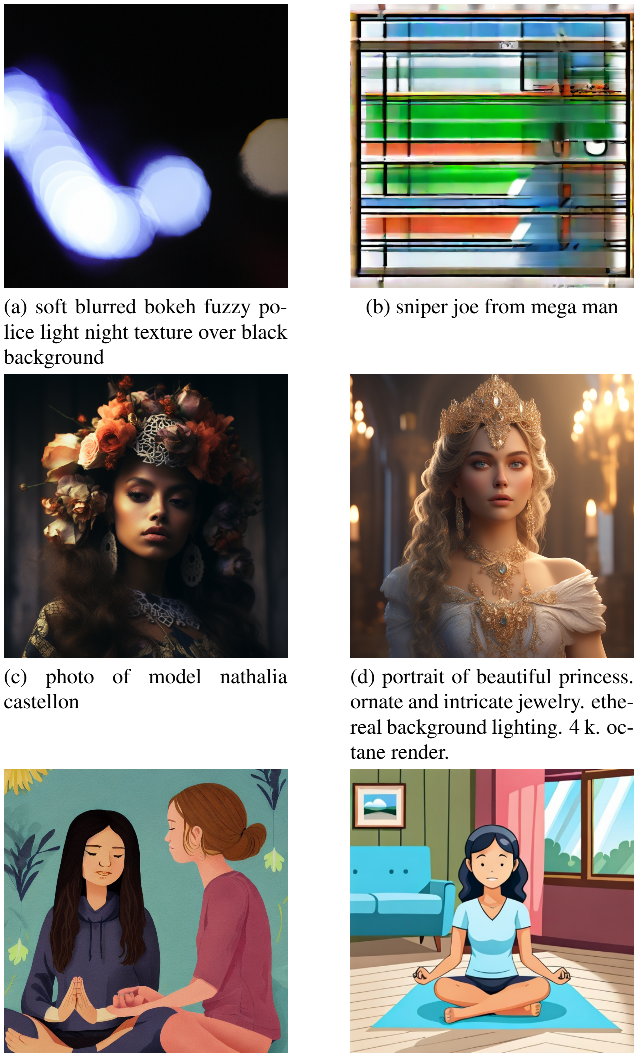
With the rapid advancement of Artificial Intelligence (AI), various models have been developed for generating AI-Generated Images (AIGIs), which have played a crucial role in various sectors such as entertainment, education, and media. However, as illustrated in Fig. 1, AIGIs often suffer from poor content quality and low similarity with the corresponding textual prompts. To guide the generation of AIGIs and evaluate the performance of AI Generated Content (AIGC) models, it is imperative to have an objective method to measure the quality of these generated images.

In the quality evaluation of AIGIs, it is essential to consider not only the degradation but also the semantic consistency between the image and the textual prompt. Typically, traditional image quality usually relies on user experience for different degradations [6, 14, 19, 27, 28, 30–32, 58, 63]. However, despite the great perceptual quality, the AIGC images inevitably suffer from unmatched textures/contents with corresponding prompts, resulting in sub-optimal generation. Semantic Similarity, denoting the alignment degree of the generated image and the provided text prompt from the perspectives of global high-level semantics, local instance semantics, and certain object attribute consistencies [17, 37, 62, 65], is another necessary aspect to measure the quality of AIGIs [13, 15, 51, 52]. The former usually relies on features extracted from pre-trained models to compute the distance [8, 44] between generated images and natural images, or fine-tuning training based on existing non-reference image quality assessment (NR-IQA) metrics [49]. The latter either employs pre-trained vision-language models [37] to extract image-text similarity or finetunes existing vision-language models [13] on extensive AIGC image quality assessment datasets.

Obtaining an AIGI score that considers both image quality and image-text similarity from these features presents a significant challenge. This score would provide a more comprehensive evaluation of the AIGI, taking into account both the visual quality of the image and its relevance to the corresponding textual prompt. Developing such a scoring system is a key research direction in the field of AI-generated content.

Therefore, we introduce SF-IQA, an advanced AIGC image quality metric that synergistically combines quality and similarity via an innovative score fusion approach. In response to the distinctive attributes of local and global distortion scopes, we utilize a multi-layered feature extraction framework to fuse local and global-level features, enabling the extraction of quality-aware representation. To obtain image-text similarity in semantic space tailored for AIGC

[*] Corresponding Author.



(a) soft blurred bokeh fuzzy police light night texture over black background

(b) sniper joe from mega man

(c) photo of model nathalia castellon

(d) portrait of beautiful princess. ornate and intricate jewelry. ethereal background lighting. 4 k. octane render.

(e) Illustration Amanda a 14-year-old girl practicing meditation and mindfulness with her mother

Figure 1. Images generated by different models and prompts. Fig. 1a and Fig. 1b have lower clarity and exhibit distortion. Fig. 1c and Fig. 1d possess rich textures and exhibit higher quality. Fig. 1e is generated from the same prompt, both reflecting “Illustration Amanda, a 14-year-old girl practicing meditation and mindfulness”. However, they exhibit noticeable differences in similarity. The image on the right lacks the “with her mother” part, resulting in a lower similarity compared to the image on the left.

image quality assessment, we finetune the vision-language model, leveraging a powerful image-text alignment prior from pic-a-pic [13]. Subsequently, the aforementioned decomposed quality components are aggregated to represent the overall quality of the AIGC image.

The contributions of this framework are summarized as follows:

- We introduce SF-IQA, a novel AIGC image quality metric that uniquely integrates image quality and text-image similarity, marking a significant advancement in the evaluation of AI-generated images.
- SF-IQA employs a sophisticated multilayer feature extraction and fusion module, effectively combining local and global-level features for accurate quality-aware feature extraction, enhancing the assessment of AIGIs.
- We introduce an innovative score fusion approach, integrating perceptual image scores with refined image-text alignment, which significantly improves AIGI quality assessment, obtaining state-of-the-art results on AGIQA-3K benchmarks and achieving 4th place in the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge [24].

2. Related work

2.1. Image Quality Assessment

No-Reference IQA (NR-IQA), which evaluates image quality without reference images, presents a more complex yet universally applicable approach [6, 14, 18, 21–23, 31, 32, 63]. In the field of NR-IQA, the two most common model architectures are CNN-based and Transformer-based models.

CNN-based IQA. Owing to the powerful feature extraction capability of CNNs, many studies have employed CNN-based models for quality evaluation. The mainstream approach involves using feature learning and regression models to learn features that adequately represent quality, thereby better-predicting image quality. CNNIQA [11] was among the first to utilize convolutional neural networks (CNNs) for IQA, predicting quality through score regression on the extracted features, and later advancing to multi-task learning frameworks [12]. WaDIQaM-NR [2] demonstrated joint learning of local quality and weights within a unified framework, enhancing representation and underscoring the relative importance of local quality in global quality estimation. HyperIQA [47] designed a hypernetwork connection to model the mapping from image content to perceived quality and integrated multi-scale local distortion features for better image quality representation.

Transformer-based IQA. Despite CNNs capturing the local structure of images, they struggle to capture non-local information and exhibit strong local bias. IQA heavily relies on both local and non-local features. Some works have utilized attention mechanisms to aid in global modeling. Maniqa [57], using ViT for feature extraction, applies attention across the channel and spatial dimensions to increase the interaction among different regions of images globally and locally. TReS [5] combines the capabilities of CNN and transformer, focusing on both local and non-local features, and reduces the bias generated by CNN’s local feature extraction, leading to better quality perception. DEIQT [35] acknowledges that features extracted solely by the Transformer encoder may not adequately express the relationship with quality, thus, it designed a decoder capable of extracting information from attention-panel embedding, allowing each part to learn image quality perception features from unique perspectives. Additionally, the expansion of transformers in large language models has led to a breakthrough in integrating large language models (LLMs) in NR-IQA. This integration combines natural language processing capabilities with IQA tasks. This multimodal approach has opened new avenues for interpretability and efficiency in IQA models [43, 48, 50].

2.2. AIGC Image Quality Assessment

With the rapid advancement of text-to-image (T2I) synthesis, image quality assessment within Artificial Intelligence Generated Content (AIGC) has become crucial for fostering research and technological advancements that align with human judgment for high-quality image generation. To support the progress in AIGC, numerous datasets have been introduced to aid the development of quality assessment tasks [13, 15, 16, 29, 49, 52, 56, 64]. Unlike conventional image quality assessment, AIGCIQA not only evaluates the quality of the image itself but also considers the consistency between the generated image content and the accompanying text [7]. Initial efforts[8, 44] in assessing the quality of generated images focused on measuring the distributional distance between generated and real images. Inspired by CLIP[37], subsequent works have shifted attention to the similarity between generated images and text prompts, adopting this perspective as an evaluative aspect in AIGCIQA. StairReward [15] assesses alignment quality down to the morpheme level, establishing a precise one-to-one correlation between image segments and their respective morphemes. PSCR [59] introduces a patch-sampling-based contrastive regression framework, leveraging the differences among various generated images to learn a more representative feature space. TIER [61] performs score regression by extracting features from both generated images and their corresponding textual prompts using respective text and image encoders. Simultaneously, IP-IQA [36] introduces a dual-stream architecture based on the CLIP model, utilizing an incremental pretraining strategy and employing a cross-attention-based image-prompt fusion module. This approach effectively merges visual and textual modalities, exploring the significance of the correlation between text prompts and generated images in assessing the quality of generated content.

3. Method

We propose a novel network architecture, SF-IQA (Score Fusion for Image Quality Assessment), which uniquely estimates the quality of the image and the similarity between the image and text in a distinct manner. This advanced approach enables SF-IQA to predict the quality from a more comprehensive perspective, making it more suitable for evaluating AIGIs.

In this section, we first provide an overview of the overall framework in Sec. 3.1. In Sec. 3.2, we describe the Quality Perception Branch, which is responsible for extracting the quality features of the image. In Sec. 3.3, we detail the Similarity Assessment Branch, which extracts the semantic similarity of the image and text. Finally, in Sec. 3.4, we describe the Score Fusion Module, which combines the quality score and the similarity score to provide a comprehensive evaluation of the AIGI's quality.

3.1. Overall Framework

As depicted in Fig. 2, the SF-IQA architecture is designed into a hybrid framework, which consists of a Quality Perception Branch, a Similarity Assessment Branch, and a Score Fusion Module. The Quality Perception Branch, rooted in the input image I, utilizes a multi-layer feature extractor to aggregate local and global-level features, facilitating the excavation of quality-aware features. The quality prediction can be represented as $Q(I)$. The Similarity Assessment Branch integrates the input image I with the textual prompt T and leverages an advanced vision-language model [13] to assess the similarity between visual and textual prompts, resulting in a similarity score, $S(I,T)$. Finally, the Score Fusion Module combines scores from both the Quality Perception and Similarity Assessment Branches using innovative fusion layers to produce the overall quality prediction for AIGI, expressed as $P(I,T)$.

$$P(I,T) = \mathcal{F}\left(Q(I), S(I,T)\right) \tag{1}$$

where $\mathcal{F}$ denotes Score Fusion Module.

3.2. Quality Perception Branch

As illustrated in Fig. 1, the local texture of AIGIs is essential for predicting the quality of AIGIs. We require a model capable of handling high-resolution images and capturing both the local texture details and global high-level information of the image. The Swin Transformer V2[26] is capable of processing high-resolution images and capturing local details of the image. Therefore, in our proposed framework, we adopt the Swin Transformer V2 as a quality perception branch $Q(\cdot)$ to extract powerful features for low-level details excavation. However, the Swin Transformer V2 falls slightly short in capturing global information. To address this, as shown in Fig. 2, we propose a Multilayer Feature Extractor and Fusion improvement to enhance the modeling of global features.

Specifically, we extract multi-level features from multiple layers within the Swin Transformer V2, which are subsequently aggregated through a transformer encoder to

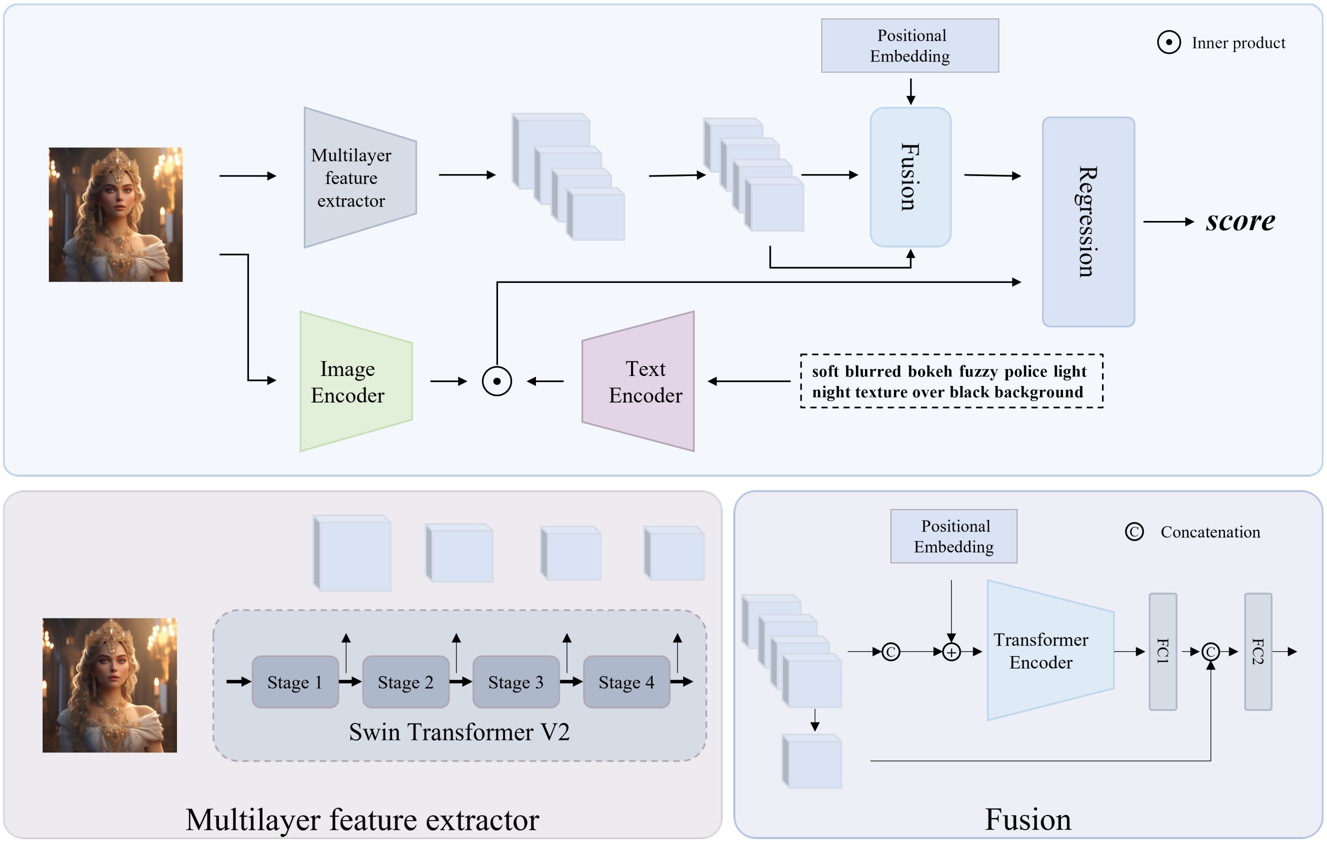


Figure 2. The proposed method's framework, wherein the Quality-aware Feature and Semantic Feature are extracted respectively by Swin Transformer V2[26] and the pre-trained CLIP-style model[13, 38]. Ultimately, the quality scores and similarity scores are fused, and they are regressed to form the MOS for the AIGIs.

obtain a global feature. The global features can provide high-level information for the quality of AIGIs. Simultaneously, the Swin Transformer V2 extracts local features from the image. These local features can capture the fine-grained details of the image. Finally, the global and local features are concatenated to construct an enhanced representation, thereby facilitating the extraction of quality-aware characteristics. This combined feature representation leverages the strengths of both global and local features, providing high-level and detailed information about the image, thereby enhancing the performance of our AIGIs quality assessment task. Finally, we fit the perceptual quality features of the image through a linear layer to obtain a predicted quality score $Q(I)$.

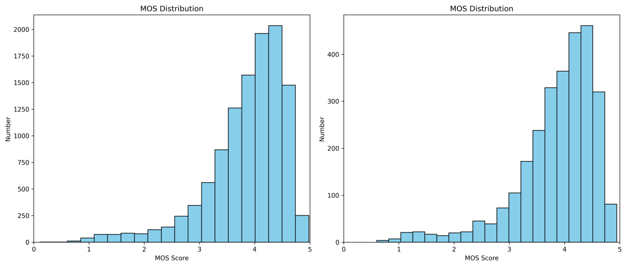
3.3. Similarity Assessment Branch

Building upon the framework established by the Contrastive Language–Image Pretraining (CLIP) model [38], we fine-tune our approach to extract and analyze semantic features from both textual prompts and their generated images. The integration of CLIP empowers our model to extract the underlying semantic similarity between text and image modalities, due to its pretraining on the web-scale image-text pairs.

Utilizing a methodology akin to the one employed in the PickScore architecture, described by [38], our similarity score function, S, intricately computes a scalar value encapsulating the degree of semantic alignment between prompt T and image I. Our scoring function is calculated using the inner product of the d-dimensional vectors I and T, produced by the image and text encoders, and scaled by a learned scalar t. The formulation of our scoring function can be succinctly expressed as:

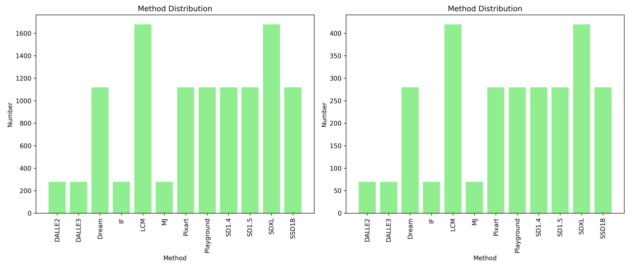
$$S(I,T) = E_{\text{txt}}(T) \cdot E_{\text{img}}(I) \cdot t \qquad (2)$$

where $E_{\text{txt}}(T)$ and $E_{\text{img}}(I)$ denote the encoded representations of the textual prompt and the image, respectively. And $S(I,T)$ quantitatively measures the similarity between the generated images and their corresponding textual descriptions.



(a) Training split (b) Testing split

Figure 3. Distribution of labeled MOS in our custom split of the AIGIQA-20K dataset.



(a) Training split (b) Testing split

Figure 4. Distribution of Method in our custom split of the AIGIQA-20K dataset.

3.4. Score Fusion Module

Upon acquiring the predicted score of quality $(Q(I)$ and similarity $S(I, T)$, we proceed by concatenating these scores into a two-dimensional vector:

$$\mathbf{v} = [Q(I), S(I, T)] \tag{3}$$

This concatenated vector, $\mathbf{v}$, is then projected into a higher-dimensional space through a fully-connected (FC) layer characterized by weights $\mathbf{W}_1$ and bias $\mathbf{b}_1$, further refined by a Rectified Linear Unit (ReLU) to introduce nonlinearity, enhancing the Score Fusion Module's capacity for complex representations. Subsequently, $\mathbf{z}$ undergoes another linear transformation, delineated by weights $\mathbf{W}_2$ and bias $\mathbf{b}_2$, to obtain a final evaluation score, $P(I, T)$. This score fulfills a comprehensive assessment of the AIGIs, encapsulating both perceptual quality and semantic similarity metrics. The process of the Score Fusion Module is encapsulated in the following equation in Eq. 4 and Eq. 5:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_1\mathbf{v} + \mathbf{b}_1) \tag{4}$$

$$P(I, T) = \mathbf{W}_2\mathbf{z} + \mathbf{b}_2 \tag{5}$$

4. Experiments

4.1. Datasets and Evaluation Criteria

Datasets. In our study, we selected four NR-IQA datasets for pre-training, including CLIVE[4], KonIQ-10K[10], LIVE[46], KADID-10K[20]. Among these, KonIQ-10K and CLIVE are real distortion datasets, while LIVE and KADID-10K are synthetic distortion datasets.

We also selected four AIGC-IQA datasets for pre-training, including AGIQA-1k[64], AGIQA-3K[15], AIGCIQA2023[49] and PKU-I2IQA[60].

For the AIGIQA-20K dataset[16], as the official validation and testing splits have not yet been released, we perform a custom train-test split on the training set and conduct training and ablation experiments on this dataset. Specifically, we divide the annotated 14000 AIGIs into two parts: 11200 AIGIs for training and 2800 AIGIs for testing. As shown in the Fig. 3 and Fig. 4, our train-test split is consistent with the distribution of MOS and AIGC methods.

Evaluation Criteria. Spearman's Rank-Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) are selected as the criteria for measuring the accuracy and consistency of the predictions, respectively. Both of these metrics range from 0 to 1. A larger SRCC indicates a more accurate ranking ability of the model, while a larger PLCC indicates a more accurate fitting ability of the model. We also use the average of PLCC and SRCC, named main score, as the performance evaluation metric for our model.

4.2. Implementation Details

Our experiments are conducted on an NVIDIA GeForce 3090 GPU, using PyTorch 2.2.0 and CUDA 11.8 for training and testing. All ablation studies are performed on the custom split of the AIGIQA-20K dataset.

Quality Perception Branch. We chose the SwinV2-T[26], pre-trained on ImageNet-1K[42], as the backbone network. For the NR-IQA datasets, we adjusted the larger edge of the image to 512 while maintaining the aspect ratio. During training, images were randomly cropped to a size of 256 x 256 and were horizontally and vertically flipped with a probability of 0.5. We used an Adam optimizer with a learning rate of 0.00002 and a weight decay of 0.0005, along with a cosine annealing scheduler. The batch size was set to 32, and the training was conducted for 70 epochs. Subsequently, we pre-trained on the AIGC-IQA datasets, using SAMA[25] to process the images to a size of 256 x 256. SAMA is an image and video processing strategy based on scaling and masking. We used an Adam optimizer with a learning rate of 0.000002 and a weight decay of 0.0005, along with a cosine annealing scheduler. The batch size was set to 32, and the training was conducted for 20 epochs.

Similarity Assessment Branch. We used PickScore[13],

Type	Metric	All		Bad Model		Medium Model		Good Model	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Hand-crafted	CEIQ[55]	0.3228	0.4166	0.1754	0.2094	0.2775	0.3043	0.1743	0.1643
	DSIQA[3]	0.4955	0.5488	0.1908	0.3139	0.2140	0.3655	0.1665	0.2520
	NIQE[32]	0.5623	0.5171	0.2031	0.3309	0.2259	0.2526	0.1750	0.2533
	Sisblim[6]	0.5479	0.6477	0.2887	0.3341	0.0540	0.2932	0.0417	0.2110
Loss-function	FID[8]	0.1733	0.1860	0.1836	0.1938	0.1402	0.1614	0.0562	0.0798
	ICS[45]	0.0931	0.0964	0.0243	0.1692	0.0797	0.1693	0.0856	0.1042
	KID[1]	0.1023	0.0786	0.0028	0.0187	0.1279	0.0860	0.0704	0.0614
SVR-based	BMPRI[31]	0.6794	0.7912	0.3686	0.4076	0.2374	0.3760	0.2046	0.2212
	GMLF[54]	0.6987	0.8181	0.3942	0.4798	0.2578	0.4036	0.0018	0.0834
	Higrade[14]	0.6171	0.7056	0.3017	0.3001	0.2376	0.2861	0.2020	0.2164
DL-based	DBCNN[63]	0.8207	0.8759	**0.5520**	**0.6825**	0.5011	0.5575	0.4288	0.4853
	CLIPIQA[48]	0.8426	0.8053	0.1882	0.2549	0.6537	0.6014	0.5038	0.5081
	CNNIQA[11]	0.7478	0.8469	0.3233	0.4547	0.4278	0.4534	0.3952	0.4517
	HyperNet[47]	0.8355	0.8903	0.5086	0.5985	0.4687	0.5480	0.5562	0.6149
	SF-IQA(quality)	**0.9024**	**0.9314**	0.4936	0.5237	**0.6739**	**0.7468**	**0.8239**	**0.8634**
Similarity	CLIP[37]	0.5972	0.6839	0.5463	0.5355	0.2272	0.2916	0.2420	0.3342
	ImageReward[52]	0.7298	0.7862	**0.5652**	0.6869	0.4464	0.5109	0.3925	0.4966
	HPS[51]	0.6349	0.7000	0.5255	0.5803	0.2762	0.3516	0.3126	0.3498
	PickScore[13]	0.6977	0.7633	0.4293	0.5588	0.3962	0.3924	0.4183	0.4743
	StairReward[15]	0.7472	0.8529	0.5401	**0.7076**	0.4642	0.5423	0.4411	0.5581
	SF-IQA(similarity)	**0.8454**	**0.9072**	0.3906	0.4345	**0.6574**	**0.7472**	**0.7144**	**0.6965**

Table 1. Performance results of quality metrics and similarity metrics on the AGIQA-3K [15] and different subsets of Text-to-Image AIGI models. The best performance results for quality prediction are marked in **red**, while the best performance results for similarity prediction are marked in **blue**.

pre-trained on Pick-a-Pic[13], as the backbone network. It is a CLIP-style model with a variant of InstructGPT's reward model objective[34].

4.3. Comparison with SOTA Results

In Tab. 1, AGIQA-3K [15] is used to conduct a comprehensive evaluation against the state-of-the-art (SOTA) metrics, we follow the dataset partitioning guidelines specified within. This methodology involved a random division of AGIQA-3K into training and testing subsets at an 80/20 ratio, with an emphasis on ensuring the image with the same object label falls into the same set.

Same as AGIQA-3K, In our analysis, AIGI models were classified into three distinct categories based on their subjective performance/alignment score, namely bad model (AttnGAN[53], GLIDE[33]), medium model (DALLE2[39], Stable Diffusion[40]), and good model (Midjourney[9], Stable Diffusion XL[41]). This classification allowed for a targeted analysis of the assessment consistency of the perception model.

With different training objectives, we developed two different models, named SF-IQA(quality) and SF-IQA(similarity). Each model was meticulously trained on its respective label type, thereby enabling an independent examination of both perceptual quality and semantic similarity.

This bespoke approach yielded superior results for both SF-IQA variants across medium and good AIGI models, as benchmarked on AGIQA-3K, establishing new SOTA metrics. Notably, the observed decline in performance metrics for models classified as bad indicates the challenges for our model inherent in extracting meaningful features from images that markedly deviate from naturalistic representations.

4.4. Ablation Studies

In order to further analyze the effectiveness of each branch and the impact of different fusion methods, we conducted ablation experiments on the custom split AIGIQA-20K dataset.

Ablation on the Image Semantic Feature Extraction. For the extraction of image semantic features (ISFE), we considered two approaches, namely detached ISFE and independent ISFE. Detached ISFE involves separating the image semantic features from the image features, which are obtained from the quality perception branch. Independent ISFE involves using an image encoder to obtain image semantic features directly from the image. The advantage

Quality	Similarity	Score Fusion	PLCC	SRCC	Main Score
✓	✓	✓	**0.9288**	**0.9006**	**0.9145**
✓	×	×	0.8931	0.8503	0.8717
×	✓	×	0.9118	0.8784	0.8951
✓	✓	×	0.9192	0.8870	0.9031

Table 2. Ablation on the branch structure and score fusion.

of the former is that it can use a simpler model structure. However, since the features obtained from the image feature extraction module also need to be used for the extraction of quality-aware features of the image, the separation of semantic features may negatively impact the extraction of quality-aware features.

The results in Tab. 3 show the effectiveness of independent ISFE(Image Semantic Feature Extraction). The practice of separating quality-aware features and semantic features from the extracted image features led to poorer results. This may be due to the separation of semantic features interfering with the extraction of quality-aware features, and the semantic features extracted in this way lack the prior for similarity prediction.

Ablation on the Similarity Score. For the similarity score, consider the use of a cross-attention module or inner product. In the Cross Attention module, we have an image semantic feature $E_{\text{img}}(I)$ as query and a text semantic feature $E_{\text{txt}}(T)$ as both key and value. We can then use the resulting attention output as a representation of the interaction between the two features to fit the final similarity score through the linear layer. The results in Tab. 4 show that the Inner Product approach is better than the Cross Attention approach. This may be due to the complexity of the structure which increases the difficulty of the model training.

Ablation on the Branch Structure. The results in Tab. 2 demonstrate the validity of the quality prediction branch and the similarity assessment branch we used. Note that the Similarity branch is actually a fine-tuned PickScore. The results show that the individual branches still have good performance in the evaluation of AIGIs, indicating that the design of each branch is reasonable and that the evaluation of AIGIs is related to quality prediction and similarity prediction respectively.

Ablation on the Score Fusion. The results in Tab. 2 prove the effectiveness of the score fusion strategy and our score fusion module. Compared with the single branch structure, using a mean of quality score and similarity score can bring about performance improvement, which indicates that the score fusion strategy is better than considering the quality scores or similarity scores alone for the evaluation of AIGIs. However, the results in Tab. 2 indicate that using a score fusion network to perform comprehensive quality fitting will bring more effective performance.

Independent ISFE	PLCC	SRCC	Main Score
×	0.8793	0.8297	0.8545
✓	**0.9288**	**0.9006**	**0.9145**

Table 3. Ablation on the image semantic feature extraction.

Similarity	PLCC	SRCC	Main Score
Inner Product	**0.9288**	**0.9006**	**0.9145**
Cross Attention	0.8861	0.8449	0.8555

Table 4. Ablation on the similarity score.

Ranking	Team name	Main Score
1	pengfei	0.9175
2	MediaSecurity_SYSU&Alibaba	0.9169
3	geniuswwg	0.9157
4	**Ours**	**0.9138**
5	QA-FTE	0.9091
6	HUTB-IQALab	0.9087
7	IQ Analyzers	0.9065
8	PKUMMCAL	0.9044
9	BDVQAGroup	0.9023
10	JNU_620	0.8835
11	MT-AIGCQA	0.8736
12	IVL	0.8715
13	z6	0.8628
14	Oblivion	0.8613
15	IVP-Lab	0.8595

Table 5. Compared to others, we get fourth place in the main score, which was obtained on the AIGIQA-20K[16] test set.

4.5. Comparison with Others

The final ranking of the test phase in NTIRE 2024 Quality Assessment of AI-Generated Content Challenge is shown in Tab. 5. The SF-IQA achieved fourth place in the main score.

5. Conclusion

In this work, we introduce a framework, SF-IQA (Score Fusion for Image Quality Assessment), which evaluates the quality of AI-Generated Images (AIGIs) by integrating image quality and image-text similarity. Structurally, SF-IQA comprises two branches that consider the quality score of the image and image-text similarity score respectively, and uses the design of the Score Fusion Module effectively to combine image quality and image-text similarity. This fusion allows for a more comprehensive and accurate assessment of AIGI quality, taking into account both the visual quality of the image and its relevance to the corresponding textual prompt.

Furthermore, experimental results in s on the AGIQA-3K database and different subsets of Text-to-Image AIGI models prove the prediction ability of SF-IQA, and ablation studies were conducted to verify the effectiveness of each component in the SF-IQA framework.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[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, 2018. 2

[3] Rony Ferzli and Lina J Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE transactions on image processing*, 18(4): 717–728, 2009. 6

[4] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 5

[5] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 2

[6] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, 60 (3):555–567, 2014. 1, 2, 6

[7] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Timo Ropinski, et al. Evaluating text to image synthesis: Survey and taxonomy of image quality metrics. *arXiv preprint arXiv:2403.11821*, 2024. 3

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 3, 6

[9] David Holz. Midjourney, 2023. https://www.midjourney.com/. 6

[10] 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

[11] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 2, 6

[12] Le Kang, Peng Ye, Yi Li, and David Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2791–2795, 2015. 2

[13] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 5, 6

[14] Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. Large-scale crowdsourced study for tone-mapped hdr pictures. *IEEE Transactions on Image Processing*, 26(10):4725–4740, 2017. 1, 2, 6

[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. 1, 3, 5, 6

[16] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Aigiqa-20k: A large database for ai-generated image quality assessment, 2024. 3, 5, 7

[17] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *arXiv preprint arXiv:2309.13625*, 2023. 1

[18] Xin Li, Yiting Lu, and Zhibo Chen. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 2023. 2

[19] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 1

[20] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 5

[21] Jianzhao Liu, Xin Li, Shukun An, and Zhibo Chen. Source-free unsupervised domain adaptation for blind image quality assessment. *arXiv preprint arXiv:2207.08124*, 2022. 2

[22] Jianzhao Liu, Xin Li, Yanding Peng, Tao Yu, and Zhibo Chen. Swiniqa: Learned swin distance for compressed image quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1795–1799, 2022.

[23] Jianzhao Liu, Wei Zhou, Xin Li, Jiahua Xu, and Zhibo Chen. Liqa: Lifelong blind image quality assessment. *IEEE Transactions on Multimedia*, 2022. 2

[24] 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. 2

[25] Yongxu Liu, Yinghui Quan, Guoyao Xiao, Aobo Li, and Jinjian Wu. Scaling and masking: A new paradigm of data sampling for image and video quality assessment. *arXiv preprint arXiv:2401.02614*, 2024. 5

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3, 4, 5

[27] Yiting Lu, Jun Fu, Xin Li, Wei Zhou, Sen Liu, Xinxin Zhang, Wei Wu, Congfu Jia, Ying Liu, and Zhibo Chen. Rtn: Reinforced transformer network for coronary ct angiography vessel-level image quality assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 644–653. Springer, 2022. 1

[28] Yiting Lu, Xin Li, Jianzhao Liu, and Zhibo Chen. Styleam: Perception-oriented unsupervised domain adaption for non-reference image quality assessment. *arXiv preprint arXiv:2207.14489*, 2022. 1

[29] Yiting Lu, Xin Li, Bingchen Li, Zihao Yu, Fengbin Guan, Xinrui Wang, Ruling Liao, Yan Ye, and Zhibo Chen. Aigcvqa: A holistic perception metric for aigc video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 3

[30] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kaleidoscope video quality assessment for short-form videos. *arXiv preprint arXiv:2402.07220*, 2024. 1

[31] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, 64 (2):508–517, 2018. 2, 6

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1, 2, 6

[33] 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. 6

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 6

[35] Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. *arXiv preprint arXiv:2304.04952*, 2023. 2

[36] Bowen Qu, Haohui Li, and Wei Gao. Bringing textual prompt to ai-generated image quality assessment. *arXiv preprint arXiv:2403.18714*, 2024. 3

[37] 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. 1, 3, 6

[38] 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. 4

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022. 6

[40] 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. 6

[41] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 6

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 5

[43] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 3

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 3

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[46] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5

[47] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3664–3673, 2020. 2, 6

[48] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence, pages 2555–2563, 2023. 3, 6

[49] 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, 3, 5

[50] 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. 3

[51] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023. 1, 6

[52] 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. 1, 3, 6

[53] 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. 6

[54] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014. 6

[55] Jia Yan, Jie Li, and Xin Fu. No-reference quality assessment of contrast-distorted images using contrast enhancement. *arXiv preprint arXiv:1904.08879*, 2019. 6

[56] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Aigcoiqa2024: Perceptual quality assessment of ai generated omnidirectional images, 2024. 3

[57] 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

[58] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Video quality assessment based on swin transformerv2 and coarse to fine strategy. *arXiv preprint arXiv:2401.08522*, 2024. 1

[59] 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. 3

[60] 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. 5

[61] Jiquan Yuan, Xinyan Cao, Jinming Che, Qinyuan Wang, Sen Liang, Wei Ren, Jinlong Lin, and Xixin Cao. Tier: Text-image encoder-based regression for aigc image quality assessment, 2024. 3

[62] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1

[63] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 1, 2, 6

[64] 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. 3, 5

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1